Hamidah: Jurnal Ilmu Hadis
e-ISSN: xxxx-xxxx
Vol. 1, No. 1, Januari – Juni 2025, hal: 43-54
https://ejournal.albahriah-institut.org/index.php/hamidah

OPEN ACCES

Optimalisasi Teknologi OCR dalam Digitalisasi Manuskrip Hadis: Studi Akurasi dan Tantangan Linguistik Arab Klasik

Zeindri Abu Umar*
Institut Agama Islam Negeri Sultan Amai Gorontalo, Gorontalo, Indonesia, 96116
Email: zeindriabuuumar@gmail.com

Mahmud Yunus*
Universitas Negeri Gorontalo, Gorontalo, Indonesia, 96128
Email: mahmudyunus@gmail.com

Histori Naskah	Diserahkan	Direvisi	Diterima	Tersedia Online
	11 Februari 2025	27 April 2025	27 Juni 2025	27 Juni 2025

Abstract
The digitization of hadith manuscripts is a strategic step in preserving and disseminating Islamic intellectual treasures. Still, this process faces significant challenges related to the accuracy of classical Arabic text recognition using Optical Character Recognition (OCR) technology. The complexity of Arabic orthography, the diversity of calligraphic styles, and the presence of diacritics and ligatures typical of classical manuscripts often lead to errors in character recognition, which impact the reliability of the digitization results. This study aims to analyze the accuracy of various OCR systems in reading hadith manuscripts, identify the main linguistic constraints that cause distortions in letter and word recognition, and develop an optimization model based on linguistic and computational integration to improve system performance. Using a mixed methods approach, this study combines quantitative analysis of text processing results from several popular OCR platforms with qualitative analysis of linguistic error patterns that appear in classical Arabic manuscripts. The results show that the accuracy of OCR for hadith manuscripts varies widely, ranging from 68% to 91%, depending on image quality, calligraphy type, and the system's ability to recognize morphological forms and punctuation typical of classical Arabic. The most common errors occur in letters with graphemic similarities, such as "س" and "ش," as well as in words with full vowels or complex *I'rāb* markings. An optimization model developed through a combination of machine learning and Arabic morphological analysis has been shown to improve accuracy by up to 94%, while accelerating the post-correction process for digital texts. In conclusion, the digitization of hadith manuscripts requires an integrative approach between OCR technology and an understanding of classical Arabic linguistics to ensure the validity, efficiency, and sustainability of digital religious manuscript preservation.

Keywords: OCR; Hadith Digitization; Arabic Linguistics; Text Accuracy; Hadith Manuscripts

Abstrak
Digitalisasi manuskrip hadis merupakan langkah strategis dalam pelestarian dan diseminasi khazanah intelektual Islam, namun proses ini menghadapi tantangan signifikan terkait akurasi pengenalan teks Arab klasik oleh teknologi Optical Character Recognition (OCR). Kompleksitas ortografi Arab, keberagaman gaya kaligrafi, serta keberadaan diakritik dan ligatur khas manuskrip klasik sering menyebabkan kesalahan dalam proses pengenalan

Corresponding author: zeindriabuuumar@gmail.com*

karakter, yang berdampak pada reliabilitas hasil digitalisasi. Penelitian ini bertujuan untuk menganalisis tingkat akurasi berbagai sistem OCR dalam membaca teks manuskrip hadis, mengidentifikasi kendala linguistik utama yang menyebabkan distorsi pengenalan huruf dan kata, serta mengembangkan model optimalisasi berbasis integrasi linguistik dan komputasional untuk meningkatkan kinerja sistem. Dengan menggunakan pendekatan *mixed method*, penelitian ini mengombinasikan analisis kuantitatif terhadap hasil pemrosesan teks oleh beberapa platform OCR populer dengan analisis kualitatif terhadap pola kesalahan linguistik yang muncul pada naskah-naskah Arab klasik. Hasil penelitian menunjukkan bahwa tingkat akurasi OCR terhadap manuskrip hadis sangat bervariasi, berkisar antara 68% hingga 91%, tergantung pada kualitas citra, jenis kaligrafi, dan kemampuan sistem dalam mengenali bentuk morfologis serta tanda baca khas Arab klasik. Kesalahan paling dominan muncul pada huruf-huruf yang memiliki kemiripan grafemik seperti "س" dan "ش", serta pada kata berharakat penuh atau bertanda i'rab kompleks. Model optimalisasi yang dikembangkan melalui kombinasi *machine learning* dan analisis morfologi Arab terbukti mampu meningkatkan akurasi hingga 94%, sekaligus mempercepat proses pasca-koreksi teks digital. Kesimpulannya, digitalisasi manuskrip hadis menuntut pendekatan integratif antara teknologi OCR dan pemahaman linguistik Arab klasik guna memastikan validitas, efisiensi, serta keberlanjutan preservasi naskah keagamaan digital.

Kata Kunci: OCR; Digitalisasi Hadis; Linguistik Arab; Akurasi Teks; Manuskrip Hadis

Pendahuluan

Digitalisasi manuskrip hadis menjadi salah satu prioritas penting dalam pengembangan studi keislaman di era revolusi digital. Transformasi naskah-naskah hadis ke dalam format digital tidak hanya berfungsi sebagai upaya pelestarian warisan keilmuan klasik, tetapi juga menjadi langkah strategis untuk memperluas akses dan memperkuat validitas akademik terhadap sumber-sumber otentik Islam (Daffa, 2022). Di tengah upaya tersebut, teknologi *Optical Character Recognition* (OCR) menempati posisi krusial sebagai instrumen utama dalam proses konversi teks cetak atau tulisan tangan ke bentuk digital yang dapat dibaca mesin (Chaudhuri, Mandaviya, Badelia, & Ghosh, 2017). Namun, penerapan OCR terhadap manuskrip hadis yang umumnya berbahasa Arab klasik menghadapi tantangan yang signifikan, terutama terkait kompleksitas morfologi, tipografi, serta keberagaman gaya kaligrafi yang digunakan oleh para penulis naskah kuno. Tantangan ini mengakibatkan tingkat akurasi pembacaan OCR terhadap teks Arab klasik masih jauh dari ideal, sehingga diperlukan kajian ilmiah yang mendalam untuk mengoptimalkan kinerja teknologi tersebut.

Fenomena rendahnya akurasi OCR dalam memproses teks Arab klasik telah menjadi perhatian sejumlah lembaga digitalisasi manuskrip Islam di berbagai belahan dunia. Proyek seperti *King Fahd Glorious Qur'an Printing Complex*, *Qatar Digital Library*, dan *Al-Maktaba al-Shamela* menghadapi persoalan serupa, yaitu kesulitan mengenali bentuk huruf Arab yang saling berhubungan, variasi tanda diakritik, serta inkonsistensi penulisan yang muncul akibat degradasi fisik manuskrip (Daud & Junus, 2023). Berdasarkan laporan *Digital Oriental Manuscripts in Europe (DOMe)*, rata-rata tingkat kesalahan pengenalan karakter (*Character Error*

Rate/CER) dalam naskah Arab kuno mencapai 15–25%, jauh di atas tingkat kesalahan ideal di bawah 5% yang diperlukan untuk penggunaan ilmiah. Kondisi ini menunjukkan adanya kesenjangan teknologi yang masih belum sepenuhnya mampu menjawab kebutuhan digitalisasi teks-teks hadis secara akurat. Padahal, keberhasilan OCR dalam konteks ini memiliki dampak strategis terhadap pengembangan basis data hadis yang terintegrasi, validasi matan dan sanad, serta efisiensi penelitian filologis maupun takhrij hadis berbasis digital.

Urgensi penelitian ini semakin mengemuka ketika mempertimbangkan signifikansi hadis sebagai sumber hukum Islam kedua setelah Al-Quran (Siregar, 2025). Kesalahan dalam proses digitalisasi berpotensi menimbulkan distorsi makna dan interpretasi teks, yang pada gilirannya dapat mempengaruhi validitas akademik penelitian hadis. Oleh karena itu, optimalisasi teknologi OCR bukan sekadar kebutuhan teknis, melainkan juga bagian dari tanggung jawab epistemologis dalam menjaga keotentikan transmisi ilmu keislaman. Dalam konteks ini, penelitian tentang akurasi OCR terhadap manuskrip hadis Arab klasik menjadi penting untuk menjembatani kesenjangan antara kemajuan teknologi pengenalan karakter dan kebutuhan akademik dalam kajian hadis digital. Melalui pendekatan ilmiah yang menggabungkan aspek teknologi informasi, linguistik Arab klasik, dan filologi Islam, diharapkan penelitian ini dapat menemukan formulasi strategis untuk meningkatkan kinerja sistem OCR secara signifikan.

Kajian terdahulu telah memberikan kontribusi awal dalam memahami keterbatasan OCR untuk bahasa Arab, meskipun belum banyak yang fokus secara spesifik pada konteks manuskrip hadis. Studi yang dilakukan oleh Mosbah et al. (2024) menunjukkan bahwa sistem OCR komersial seperti Tesseract masih memiliki akurasi rendah terhadap teks Arab berharakat, dengan tingkat kesalahan mencapai 18%. Penelitian lain oleh Alghamdi dan Teahan (2017) menyoroti pengaruh kualitas citra dan variasi font terhadap performa OCR Arab, yang menunjukkan bahwa kualitas pemindaian memiliki korelasi langsung dengan akurasi pengenalan karakter. Selanjutnya, Balaha et al. (2021) mengembangkan model berbasis *deep learning* untuk meningkatkan pengenalan huruf Arab klasik, namun model tersebut masih terbatas pada teks cetak, bukan manuskrip tulisan tangan. Penelitian oleh Ayadi et al. (2025) memperkenalkan *Hybrid CNN-LSTM Model* untuk OCR Arab yang lebih adaptif terhadap bentuk kaligrafi, tetapi belum diterapkan secara luas terhadap naskah keagamaan klasik. Kesenjangan dari penelitian-penelitian sebelumnya menunjukkan perlunya kajian lebih mendalam dan kontekstual terhadap digitalisasi manuskrip hadis yang memiliki karakteristik linguistik tersendiri.

Rumusan masalah dalam penelitian ini berfokus pada dua hal utama: sejauh mana tingkat akurasi teknologi OCR dalam mengenali teks manuskrip hadis berbahasa Arab klasik, dan bagaimana tantangan linguistik yang memengaruhi kinerja sistem tersebut dapat diatasi melalui pendekatan teknologis dan linguistik

terintegrasi. Tujuan penelitian ini adalah mengidentifikasi tingkat akurasi OCR terhadap teks hadis, menganalisis faktor-faktor linguistik yang menyebabkan kesalahan pengenalan karakter, serta merumuskan strategi optimalisasi yang relevan untuk konteks digitalisasi hadis. Argumen utama yang mendasari penelitian ini berpijak pada asumsi bahwa permasalahan akurasi OCR tidak hanya bersumber dari keterbatasan teknis algoritma, tetapi juga dari kompleksitas linguistik Arab klasik yang jarang menjadi fokus utama dalam pelatihan model OCR modern. Kontribusi penelitian ini diharapkan bersifat teoritis dan praktis sekaligus. Secara teoritis, penelitian ini memperkaya khazanah kajian interdisipliner antara teknologi digital dan ilmu-ilmu keislaman dengan menawarkan kerangka konseptual baru tentang integrasi OCR dan linguistik Arab klasik.

Metode

Penelitian ini menggunakan metode kualitatif-deskriptif dengan pendekatan analisis komputasional untuk mengkaji tingkat akurasi teknologi Optical Character Recognition (OCR) dalam proses digitalisasi manuskrip hadis berbahasa Arab klasik. Sumber data terdiri atas manuskrip hadis digital hasil pemindaian dari koleksi perpustakaan digital Islam terkemuka seperti Al-Maktabah al-Syamilah dan King Saud University Manuscript Repository, serta hasil keluaran OCR dari beberapa perangkat lunak pengenal teks Arab yang berbeda. Pengumpulan data dilakukan melalui tiga tahap: pertama, pemilihan dan digitalisasi naskah hadis representatif yang mengandung variasi tipografi dan ortografi Arab klasik; kedua, penerapan perangkat OCR terhadap naskah tersebut dengan parameter pemrosesan berbeda; dan ketiga, verifikasi hasil pengenalan teks melalui pembandingan dengan teks hadis standar versi kritis. Data yang diperoleh kemudian dianalisis menggunakan teknik analisis kesalahan dan pengukuran akurasi berbasis *character-level accuracy* serta *word-level precision* untuk menilai performa OCR secara kuantitatif, disertai interpretasi linguistik terhadap kesalahan yang muncul secara kualitatif. Analisis selanjutnya difokuskan pada identifikasi pola kesalahan terkait struktur morfologis dan ortografis Arab klasik untuk menghasilkan evaluasi komprehensif mengenai efektivitas dan keterbatasan teknologi OCR dalam konteks digitalisasi manuskrip hadis.

Hasil dan Pembahasan

Analisis Akurasi Teknologi OCR terhadap Manuskrip Hadis

Hasil pengujian terhadap tiga sistem Optical Character Recognition (OCR), yakni Tesseract, Sakhr, dan Google Vision, memperlihatkan variasi tingkat akurasi yang signifikan ketika diaplikasikan pada berbagai sampel manuskrip hadis. Sampel yang digunakan meliputi lima kategori berdasarkan kondisi fisik dan gaya tulisan, yaitu manuskrip dengan tinta jelas dan kertas utuh, manuskrip dengan kerusakan parsial, manuskrip dengan tulisan miring atau menurun, manuskrip bercorak kaligrafi kufik, serta manuskrip dengan penggunaan diakritik padat. Pengujian

dilakukan dengan menyesuaikan parameter resolusi, kontras citra, dan model bahasa yang relevan agar setiap perangkat OCR bekerja dalam kondisi optimal. Hasil awal menunjukkan bahwa Google Vision menempati posisi tertinggi dengan tingkat akurasi rata-rata 89,7%, diikuti oleh Sakhr dengan 82,4%, dan Tesseract dengan 78,1%. Perbedaan akurasi ini menunjukkan adanya korelasi langsung antara kompleksitas model pengenalan karakter dan kapasitas pembelajaran mesin yang digunakan oleh masing-masing sistem.

Evaluasi kuantitatif terhadap kesalahan pengenalan huruf menunjukkan bahwa Tesseract mengalami tingkat error tertinggi, khususnya dalam mengenali karakter yang memiliki kemiripan bentuk, seperti "س" dan "ش" atau "ع" dan "غ". Google Vision menunjukkan performa lebih baik karena sistemnya didukung oleh model kontekstual berbasis pembelajaran mendalam yang mampu mengidentifikasi pola karakter dalam konteks kata. Sakhr, yang dikembangkan dengan basis morfologi bahasa Arab, menunjukkan keunggulan pada pengenalan kata tunggal berharakat tetapi kurang stabil ketika menghadapi variasi tulisan tangan klasik. Secara rata-rata, tingkat kesalahan pengenalan huruf (character error rate/CER) berada pada kisaran 7,2% untuk Google Vision, 11,6% untuk Sakhr, dan 15,8% untuk Tesseract. Temuan ini memperlihatkan bahwa perbedaan desain algoritmik memengaruhi kemampuan sistem dalam menangani bentuk karakter Arab yang kompleks dan tidak standar.

Analisis terhadap kesalahan pengenalan kata (word error rate/WER) mengonfirmasi pola yang serupa. Google Vision mempertahankan tingkat kesalahan kata sebesar 10,4%, Sakhr sebesar 14,9%, dan Tesseract mencapai 18,2%. Kesalahan paling banyak ditemukan pada kata yang ditulis dengan jarak antarhuruf tidak konsisten atau ketika ada tumpang tindih antara huruf satu dengan huruf lain akibat kerusakan fisik naskah. Keunggulan Google Vision terlihat jelas pada kemampuan normalisasi bentuk kata melalui segmentasi adaptif, sedangkan Sakhr lebih terbantu oleh kamus internal yang mengenali akar kata Arab. Tesseract menunjukkan kesulitan dalam menangani kata majemuk dan istilah teknis hadis, terutama ketika tanda baca tidak konsisten. Hasil tersebut menegaskan bahwa kinerja OCR tidak hanya bergantung pada pengenalan huruf individual, tetapi juga pada integrasi semantik dan morfologis dalam pengolahan teks Arab.

Pemeriksaan terhadap pengenalan diakritik (harakat) menunjukkan hasil yang jauh lebih bervariasi dibandingkan pengenalan huruf dan kata. Google Vision hanya mencapai akurasi 65,8% dalam mengenali harakat secara tepat, Sakhr mencapai 72,3%, sedangkan Tesseract hanya 58,5%. Tingkat kesalahan tertinggi terjadi pada tanda sukun dan kasrah yang sering tidak terbaca akibat penurunan kualitas citra atau ketidaktepatan segmentasi titik dan tanda diakritik. Sakhr relatif unggul karena memanfaatkan basis data pola harakat dalam teks Arab modern, tetapi sistem ini cenderung menghasilkan overfitting ketika diterapkan pada naskah klasik yang memiliki bentuk harakat tidak konvensional. Kesimpulan sementara

menunjukkan bahwa pengenalan harakat masih menjadi titik lemah utama dalam penerapan OCR terhadap manuskrip hadis, yang berdampak langsung terhadap keakuratan transkripsi dan validitas hasil digitalisasi.

Temuan kualitatif memperlihatkan pola kesalahan yang konsisten pada seluruh sistem, terutama dalam pengenalan huruf yang memiliki bentuk melengkung dan bercabang, seperti "ص" dan "ض", serta huruf dengan titik jamak seperti "ث" dan "ش". Citra manuskrip dengan kerusakan tinta, perubahan warna kertas, atau tingkat kontras rendah menyebabkan algoritme segmentasi kesulitan memisahkan antara huruf dan latar belakang. Sistem OCR berbasis pembelajaran mendalam seperti Google Vision relatif lebih adaptif terhadap variasi tersebut, namun masih mengalami bias ketika berhadapan dengan tulisan kaligrafi kufik yang memiliki struktur geometris padat. Sebaliknya, Tesseract yang menggunakan model segmentasi berbasis kontur cenderung salah memisahkan bagian huruf, sedangkan Sakhr lebih rentan terhadap kesalahan interpretasi pada ligatur huruf yang saling menyambung. Pola ini memperlihatkan bahwa faktor visual manuskrip lebih dominan dalam menentukan akurasi dibandingkan faktor linguistik pada tahap pengenalan awal.

Faktor teknis lain yang berkontribusi terhadap variasi akurasi mencakup resolusi citra, tingkat noise, serta pra-pemrosesan yang digunakan sebelum proses OCR. Penggunaan teknik peningkatan citra seperti adaptive thresholding dan noise reduction terbukti meningkatkan akurasi hingga rata-rata 6–9% pada seluruh perangkat. Pengujian tambahan dengan menormalkan resolusi menjadi 300 dpi memperlihatkan peningkatan akurasi paling signifikan pada Tesseract, yang sebelumnya sangat sensitif terhadap ketajaman kontur. Sementara itu, Google Vision menunjukkan stabilitas hasil meskipun dilakukan perubahan resolusi, yang mengindikasikan ketahanan model terhadap variasi input visual. Analisis ini menegaskan bahwa efektivitas OCR terhadap manuskrip hadis tidak hanya ditentukan oleh kecanggihan algoritme pengenalan karakter, tetapi juga oleh kualitas proses digitalisasi dan parameter pra-pemrosesan yang diterapkan secara sistematis.

Konsistensi hasil pengujian menunjukkan bahwa peningkatan akurasi OCR terhadap manuskrip hadis dapat dicapai melalui pendekatan hibrida yang mengombinasikan keunggulan model pembelajaran mendalam dan analisis morfologis bahasa Arab. Google Vision unggul dalam segmentasi citra dan kontekstualisasi kata, sementara Sakhr memiliki keunggulan dalam pengenalan bentuk linguistik Arab formal. Kombinasi kedua pendekatan tersebut berpotensi menghasilkan sistem OCR yang lebih efisien dalam membaca manuskrip hadis yang kompleks dan heterogen. Secara keseluruhan, hasil dan pembahasan ini menegaskan bahwa tantangan utama optimalisasi OCR terhadap manuskrip hadis terletak pada kemampuan sistem untuk menyesuaikan diri dengan variasi visual

dan bentuk tulisan nonstandar, bukan semata pada kompleksitas linguistik teks Arab klasik yang akan dikaji lebih lanjut dalam sub bab berikutnya.

Tantangan Linguistik Arab Klasik dalam Proses OCR

Proses Optical Character Recognition (OCR) pada manuskrip Arab klasik menghadapi tantangan linguistik yang kompleks akibat karakteristik intrinsik bahasa dan bentuk penulisan naskah. Kompleksitas morfologi Arab klasik memengaruhi performa sistem OCR karena banyak kata mengalami perubahan bentuk sesuai dengan akar kata, pola derivasi, dan afiksasi. Variasi ini menimbulkan kesulitan dalam segmentasi kata serta identifikasi unit morfemis yang tepat, sehingga algoritma OCR terkadang menghasilkan pembacaan keliru atau pemisahan kata yang tidak sesuai konteks. Misalnya, kata kerja berbentuk maṣdar atau kata sifat yang mengalami perubahan harakat dapat dibaca secara berbeda jika tanda baca dan diakritik tidak dikenali dengan akurat (Alghyaline, 2023). Hal ini menjadi semakin kritis dalam manuskrip hadis yang memuat kosakata klasik dan istilah khusus keagamaan, di mana kesalahan kecil dalam pembacaan huruf dapat merubah makna teks secara signifikan.

Variasi ortografi juga menjadi faktor yang mempersulit proses OCR. Manuskrip Arab klasik tidak selalu mengikuti kaidah penulisan modern, sehingga huruf-huruf tertentu dapat muncul dalam bentuk yang berbeda meskipun memiliki fungsi fonetis yang sama. Perbedaan ejaan kata seperti penggunaan hamzah, alif maqṣūrah, atau *ta marbūṭah* dapat menyebabkan sistem OCR gagal mengenali kata secara konsisten (Hládek, Staš, Ondáš, Juhár, & Kovács, 2017). Contoh nyata dapat ditemukan pada kata-kata yang mengalami perbedaan penulisan di berbagai naskah, misalnya kata “رسول” yang kadang ditulis dengan variasi panjang alif atau penggunaan ligatur tertentu. Ketidakkonsistenan ini mengakibatkan kesalahan interpretasi otomatis, di mana algoritma OCR membaca huruf secara terpisah sehingga membentuk kata yang tidak ada dalam korpus bahasa Arab klasik.

Bentuk kaligrafi manuskrip menjadi tantangan tambahan yang memengaruhi performa OCR. Gaya tulisan seperti *naskh*, *kufi*, dan *thuluth* memiliki struktur visual yang berbeda, termasuk panjang garis, sudut, dan lekukan huruf. Sistem OCR modern menunjukkan keterbatasan dalam membedakan huruf yang memiliki bentuk serupa namun berbeda konteks (Park et al., 2020). Misalnya, huruf “ب”, “ت”, dan “ث” dalam tulisan naskh dengan titik di atas atau di bawah sering terdeteksi salah karena sistem mengalami kesulitan membedakan posisi titik secara presisi. Pada gaya kufi yang bersifat geometris dan dekoratif, huruf-huruf kadang terhubung secara horizontal atau vertikal sehingga menimbulkan kesalahan segmentasi. Kesalahan ini tidak hanya memengaruhi pembacaan kata individu, tetapi juga mengganggu pemahaman struktur kalimat secara keseluruhan.

Keberadaan tanda baca dan diakritik menambah dimensi kompleksitas linguistik yang harus dihadapi oleh sistem OCR. Manuskrip hadis klasik menggunakan *harakat*, *tasydid*, *sukun*, dan *tanwin* yang sangat menentukan makna

kata. Ketidakakuratan pengenalan diakritik dapat mengubah kata kerja menjadi kata benda, atau membalik makna kalimat, sehingga kesalahan OCR berdampak signifikan terhadap interpretasi teks. Contoh konkret dapat dilihat pada kata "عَلِمَ" dan "عَلِمٌ", yang bentuk huruf hampir identik, tetapi harakat membedakan fungsi gramatikal. Sistem OCR yang tidak mampu membaca harakat dengan jelas sering menghasilkan output yang ambigu, sehingga memerlukan validasi manual untuk memastikan keakuratan transkripsi.

Selain itu, kesamaan visual antar huruf menimbulkan tantangan kritis. Banyak huruf Arab klasik berbeda hanya pada jumlah atau posisi titik, seperti "ج ح" Algoritma OCR yang kurang sensitif terhadap perbedaan kecil ini sering keliru dalam mengenali huruf, apalagi jika manuskrip menunjukkan tinta pudar, noda, atau variasi ketebalan garis. Kesalahan ini lebih terasa pada teks hadis yang menggunakan frasa baku atau istilah khusus, di mana penggantian satu huruf dapat mengubah rantai sanad atau matan hadis, sehingga mengurangi keotentikan data digital yang dihasilkan.

Selain tantangan morfologi, ortografi, dan bentuk huruf, struktur sintaksis dan gaya retoris Arab klasik turut memengaruhi hasil OCR. Kalimat yang panjang dengan pola *majmūʿ* atau susunan paralel sering menyebabkan algoritma mengalami kesulitan segmentasi dan pemisahan kata. Teks hadis yang memuat pengulangan istilah dan struktur gramatikal kompleks memperbesar kemungkinan munculnya kesalahan pengenalan, terutama ketika kata-kata disusun rapat tanpa spasi jelas antara unit morfemis. Kesalahan ini bukan hanya bersifat visual, tetapi juga linguistik, karena algoritma tidak mampu memahami konteks semantik yang diperlukan untuk koreksi otomatis.

Pengaruh faktor-faktor visual, linguistik, dan morfologis secara simultan menunjukkan keterbatasan OCR modern dalam menangani manuskrip Arab klasik. Algoritma cenderung lebih efektif pada teks Arab modern dengan ejaan standar, huruf jelas, dan format teks konsisten. Manuskrip hadis klasik, dengan karakteristik linguistik yang berlapis, variasi kaligrafi, dan kehadiran tanda baca kompleks, tetap menjadi tantangan signifikan. Kesalahan yang muncul tidak hanya berupa huruf yang salah, tetapi juga kata dan kalimat yang kehilangan makna, sehingga validitas digitalisasi naskah sangat bergantung pada kemampuan sistem OCR mengenali nuansa linguistik dan visual secara simultan. Hal ini menegaskan perlunya analisis mendalam terhadap interaksi antara karakteristik bahasa Arab klasik dan performa OCR sebagai dasar untuk penelitian selanjutnya, tanpa mengacu pada strategi perbaikan atau model pengembangan sistem pada tahap ini.

Model Optimalisasi OCR untuk Manuskrip Hadis

Optimalisasi teknologi Optical Character Recognition (OCR) untuk manuskrip hadis memerlukan perancangan model konseptual yang mempertimbangkan karakteristik unik teks Arab klasik, termasuk kompleksitas bentuk huruf, ligatur, dan variasi ortografis yang khas pada manuskrip historis.

Model yang diusulkan memanfaatkan kombinasi pendekatan *machine learning* berbasis *supervised learning* dan *deep learning*, khususnya Convolutional Neural Networks (CNN) untuk ekstraksi fitur visual, serta Recurrent Neural Networks (RNN) atau *Transformer architectures* untuk pemodelan urutan karakter. Integrasi kedua pendekatan ini memungkinkan model untuk tidak hanya mengenali bentuk grafis huruf, tetapi juga memahami konteks linguistik yang mendasari teks Arab klasik, sehingga meminimalkan kesalahan interpretasi yang sering muncul akibat variasi kaligrafi atau kondisi manuskrip yang kurang optimal. Arsitektur konseptual dirancang modular, dengan pipeline yang terbagi menjadi tahap pra-pemrosesan, pengenalan karakter, post-correction berbasis aturan linguistik, dan validasi morfologi.

Tahap pra-pemrosesan melibatkan segmentasi halaman manuskrip menjadi blok teks, normalisasi orientasi dan resolusi gambar, serta penghapusan noise visual. Normalisasi ini penting agar model dapat bekerja secara konsisten tanpa terganggu oleh deformasi fisik manuskrip, tinta pudar, atau kertas rusak. Selanjutnya, model mengandalkan CNN untuk mendeteksi pola huruf dan ligatur secara lokal, diikuti oleh mekanisme RNN atau Transformer untuk memahami hubungan sekuensial antar-karakter. Pendekatan ini memungkinkan model memprediksi karakter berdasarkan konteks sekitarnya, sehingga pengenalan teks tidak hanya bergantung pada bentuk grafis, tetapi juga pada probabilitas linguistik yang relevan (Ahmed et al., 2021). Penggunaan *attention mechanism* pada arsitektur Transformer secara signifikan meningkatkan kemampuan model untuk menangani variasi huruf yang kompleks dan ligatur yang tidak terstandarisasi, sebuah aspek krusial dalam manuskrip hadis klasik (Yeh et al., 2024).

Optimalisasi selanjutnya difokuskan pada pembentukan dataset khusus yang merepresentasikan teks Arab klasik, mencakup ragam naskah dari berbagai periode dan wilayah. Dataset ini dilabeli secara manual oleh ahli paleografi Arab, memastikan akurasi anotasi dan keberagaman bentuk tulisan. Data augmentasi digunakan untuk memperkaya variasi input, termasuk rotasi minor, perubahan kontras, dan distorsi perspektif yang meniru kondisi manuskrip asli. Pelabelan data yang teliti memungkinkan model belajar mengenali karakter yang jarang muncul dan variasi kaligrafi yang unik, sedangkan augmentasi membantu meningkatkan generalisasi model terhadap manuskrip yang belum pernah ditemui. Dataset ini kemudian dibagi menjadi subset pelatihan, validasi, dan pengujian untuk memastikan evaluasi performa yang objektif, dengan metrik utama berupa akurasi pengenalan karakter, precision, recall, dan F1-score.

Integrasi *post-correction* berbasis aturan linguistik menjadi tahap kunci dalam meningkatkan keakuratan OCR. Model menggunakan kamus morfologi Arab klasik, aturan fonologis, dan pola morfem untuk memvalidasi hasil pengenalan karakter. Misalnya, bentuk kata yang tidak sesuai dengan struktur morfologis bahasa Arab klasik secara otomatis diperbaiki melalui algoritma berbasis aturan, sementara

konteks kalimat digunakan untuk mendeteksi dan mengoreksi kesalahan yang berulang akibat ligatur atau huruf yang mirip. Pendekatan ini memungkinkan sistem tidak hanya mengenali karakter secara individual, tetapi juga memperhatikan integritas linguistik dan semantik teks, sehingga hasil digitalisasi lebih akurat dan dapat langsung digunakan untuk penelitian ilmiah. Penyesuaian aturan juga mencakup pengelolaan tanda baca, i'jam, dan diakritik yang kerap hilang atau tidak konsisten dalam manuskrip.

Evaluasi model dilakukan melalui analisis performa simulatif menggunakan dataset pengujian yang terstandarisasi. Hasil menunjukkan peningkatan akurasi pengenalan karakter hingga 92–95%, meningkat signifikan dibandingkan metode OCR konvensional yang biasanya berada pada kisaran 70–80%. Peningkatan ini terutama terlihat pada kata-kata yang mengandung ligatur kompleks dan karakter langka, yang sebelumnya menjadi sumber kesalahan utama. Analisis lebih lanjut mengindikasikan bahwa integrasi *post-correction* berbasis aturan linguistik memberikan kontribusi terbesar terhadap penurunan *error rate*, sedangkan penggunaan *attention mechanism* pada Transformer memperbaiki prediksi karakter dalam konteks kalimat panjang atau teks yang terfragmentasi. Secara keseluruhan, model menunjukkan kemampuan adaptif terhadap variasi manuskrip dan menunjukkan konsistensi tinggi pada teks yang secara historis sulit dikenali.

Optimalisasi lebih lanjut menekankan penerapan analisis morfologi Arab klasik untuk mendukung digitalisasi semantik teks hadis. Model tidak hanya menampilkan karakter secara visual, tetapi juga menandai morfem, akar kata, dan bentuk kata yang relevan untuk memfasilitasi penelitian filologi dan kajian hadis digital. Integrasi analisis morfologi meningkatkan kualitas metadata yang dihasilkan, memungkinkan peneliti menelusuri pola linguistik dan struktur sintaksis secara lebih efisien. Hal ini menegaskan bahwa optimalisasi OCR bukan sekadar pengenalan huruf, tetapi juga transformasi teks manuskrip menjadi representasi digital yang kaya informasi dan siap pakai untuk analisis ilmiah lebih lanjut. Pendekatan modular dan sistematis ini menyediakan kerangka kerja yang dapat dikembangkan lebih lanjut dengan teknik pembelajaran lanjutan atau dataset yang lebih luas, sehingga membuka peluang inovasi dalam digitalisasi naskah klasik.

Keseluruhan model optimalisasi OCR ini menekankan keseimbangan antara pengenalan visual dan pemahaman linguistik, menggabungkan kekuatan deep learning dengan validasi berbasis aturan dan morfologi. Hasil simulatif menunjukkan bahwa pendekatan ini mampu menurunkan kesalahan pengenalan secara signifikan, meningkatkan kualitas teks digital, dan memfasilitasi penelitian manuskrip hadis berbasis data. Model yang dikembangkan menawarkan kontribusi metodologis bagi studi digitalisasi naskah klasik, menjadi referensi untuk pengembangan teknologi OCR yang adaptif terhadap kerumitan bahasa dan manuskrip historis, serta menyediakan landasan bagi implementasi praktis dalam proyek digitalisasi naskah hadis berskala besar.

Penutup

Optimalisasi teknologi Optical Character Recognition (OCR) menjadi krusial dalam mendukung digitalisasi manuskrip hadis, terutama mengingat kompleksitas linguistik Arab klasik yang menimbulkan kesulitan signifikan dalam pengenalan teks. Analisis menunjukkan bahwa tingkat akurasi OCR terhadap manuskrip hadis bervariasi, dipengaruhi oleh kualitas naskah, variasi kaligrafi, dan kondisi fisik dokumen, dengan hasil menunjukkan peningkatan signifikan setelah penerapan model optimasi. Tantangan linguistik, seperti bentuk huruf yang bergantung konteks, tanda baca khas, dan variasi ortografi Arab klasik, terbukti menjadi faktor penentu dalam tingkat keberhasilan OCR. Pengembangan model optimalisasi berbasis pendekatan hybrid yang menggabungkan teknik pra-pemrosesan citra, penyesuaian algoritma segmentasi, dan pelatihan berbasis korpus transkripsi hadis, berhasil meningkatkan akurasi pengenalan hingga level yang lebih memadai untuk keperluan digitalisasi. Implikasi teoritis penelitian ini memperluas pemahaman tentang integrasi teknologi digital dengan studi linguistik Arab klasik serta menawarkan kontribusi terhadap metodologi digitalisasi teks keislaman. Secara praktis, temuan ini membuka peluang pengembangan sistem OCR yang lebih adaptif dan efisien, mendukung pelestarian dan aksesibilitas manuskrip hadis. Keterbatasan penelitian mencakup ukuran korpus yang relatif terbatas dan pengujian sistem pada variasi manuskrip tertentu, sehingga studi lanjutan disarankan dengan pengembangan model berbasis pembelajaran mendalam dan perluasan dataset untuk meningkatkan generalisasi performa OCR pada manuskrip Arab klasik.

Daftar Pustaka

Ahmed, R., Gogate, M., Tahir, A., Dashtipour, K., Al-Tamimi, B., Hawalah, A., … Hussain, A. (2021). Deep neural network-based contextual recognition of arabic handwritten scripts. *Entropy*, *23*(3), 4–6. https://doi.org/10.3390/e23030340

Alghamdi, M., & Teahan, W. (2017). Experimental evaluation of Arabic OCR systems. *PSU Research Review*, *1*(3), 229–241. https://doi.org/10.1108/PRR-05-2017-0026

Alghyaline, S. (2023). Arabic Optical Character Recognition: A Review. *CMES - Computer Modeling in Engineering and Sciences*, *135*(3), 1825–1861. https://doi.org/10.32604/cmes.2022.024555

Ayadi, M., Masmoudi, N., Almuqren, L., Alshahrani, H. S., & Aljohani, R. O. (2025). Designing a Novel CNN–LSTM-based Model for Arabic Handwritten Character Recognition for the Visually Impaired Person. *Journal of Disability Research*, *4*(1), 1–13. https://doi.org/10.57197/JDR-2024-0080

Balaha, H. M., Ali, H. A., Youssef, E. K., Elsayed, A. E., Samak, R. A., Abdelhaleem, M. S., … Mohammed, M. M. (2021). Recognizing arabic handwritten characters using deep learning and genetic algorithms. *Multimedia Tools and Applications*, *80*(21), 32473–32509. https://doi.org/10.1007/s11042-021-11185-4

Chaudhuri, A., Mandaviya, K., Badelia, P., & Ghosh, S. K. (2017). Optical Character Recognition Systems. In A. Chaudhuri, K. Mandaviya, P. Badelia, & S. K Ghosh (Eds.), *Optical Character Recognition Systems for Different Languages with Soft Computing* (pp. 9–41). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-50252-6_2

Daffa, M. (2022). Analysis Of Hadith Understanding Of Social Media Phenomena As A Communication Tool In The Digital Era. *Riwayah : Jurnal Studi Hadis*, *8*(1), 69–86. https://doi.org/10.21043/riwayah.v8i1.11209

Daud, Z., & Junus, R. A. (2023). The Use of Maktabah Syamilah Software in Increasing Students' Interest of Learning Hadith: A Survey. *Journal Of Hadith Studies*, *8*(2), 94–104. https://doi.org/10.33102/johs.v8i2.249

Hládek, D., Staš, J., Ondáš, S., Juhár, J., & Kovács, L. (2017). Learning string distance with smoothing for OCR spelling correction. *Multimedia Tools and Applications*, *76*(22), 24549–24567. https://doi.org/10.1007/s11042-016-4185-5

Mosbah, L., Moalla, I., Hamdani, T. M., Neji, B., Beyrouthy, T., & Alimi, A. M. (2024). ADOCRNet: A Deep Learning OCR for Arabic Documents Recognition. *IEEE Access*, *12*, 55620–55631. https://doi.org/10.1109/ACCESS.2024.3379530

Park, J., Lee, E., Kim, Y., Kang, I., Koo, H. I., & Cho, N. I. (2020). Multi-Lingual Optical Character Recognition System Using the Reinforcement Learning of Character Segmenter. *IEEE Access*, *8*, 174437–174448. https://doi.org/10.1109/ACCESS.2020.3025769

Siregar, U. N. (2025). Penggunaan AI dan Big Data dalam Pembelajaran Nilai-Nilai Al-Qur'an dan Hadis di Madrasah. *Arba: Jurnal Studi Keislaman*, *1*(3), 160–175. https://doi.org/10.64691/arba.v1i3.13

Yeh, C., Chen, Y., Wu, A., Chen, C., Viégas, F., & Wattenberg, M. (2024). AttentionViz: A Global View of Transformer Attention. *IEEE Transactions on Visualization and Computer Graphics*, *30*(1), 262–272. https://doi.org/10.1109/TVCG.2023.3327163