International Journal Business, Management and Innovation Review

e-ISSN : 3046-5605
p-ISSN :3046-6423

Research Article

A Comparative Analysis of the Accuracy and Efficiency of OCR Tools in Digital Archiving at Government Agencies in Malang City

Fahmi Muhammad Azzuhri[1*], Farika Nikmah[2], Joni Dwi Pribadi[3]

[1]Politeknik Negeri Malang; Indonesia; e-mail : fahmi.azzuhri@gmail.com
[2]Politeknik Negeri Malang; Indonesia; e-mail : farika.nikmah@polinema.ac.id
[3]Politeknik Negeri Malang; Indonesia; e-mail : joni.dwi@polinema.ac.id
*Corresponding Auhtor: fahmi.azzuhri@gmail.com

Abstract: The digitization of government archives frequently yields non-machine-readable image files, limiting searchability and administrative utility. This study comparatively evaluates the accuracy and efficiency of four Optical Character Recognition (OCR) tools (Tesseract, Google Vision, ABBYY FineReader, and Microsoft OCR) applied to administrative archives from government agencies in Malang City. Employing a comparative quantitative evaluative design, the research measures performance using Character Error Rate (CER), Word Error Rate (WER), and document processing time across purposively sampled documents. Data were analyzed via descriptive statistics and One-Way ANOVA. Results indicate that ABBYY FineReader achieves the highest accuracy (mean CER: 0.041; WER: 0.078), whereas Google Vision OCR demonstrates optimal processing efficiency (mean: 3.44 s). ANOVA confirms statistically significant performance differences across all metrics ($p < 0.05$), revealing a clear accuracy–efficiency trade-off. These findings demonstrate that no single tool universally dominates; rather, OCR selection must align with document criticality and institutional workflow demands. Consequently, high-precision archives benefit from accuracy-optimized systems, while high-volume digitization prioritizes computational speed. This research provides empirical evidence for context-driven OCR implementation, offering public institutions a structured decision framework to enhance digital archive accessibility, administrative efficiency, and long-term information management sustainability.

Keywords: Comparative Performance Evaluation; Digital Archives; Document Digitization; Government Administration; Optical Character Recognition (OCR).

Received: January 30, 2026
Revised: February 28, 2026
Accepted: April 26, 2026
Published: July 08, 2026
Curr. Ver.: July 08, 2026

Copyright: © 2025 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY SA) license (https://creativecommons.org/licenses/by-sa/4.0/)

1. Introduction

Digital transformation has fundamentally reshaped records management in the public sector. Government agencies are increasingly expected to enhance service efficiency, administrative transparency, bureaucratic accountability, and information accessibility. Within this paradigm, the digitization of archival records has emerged as a strategic imperative to overcome the limitations of conventional paper-based systems. Digital archives function not merely as storage repositories but as integral components of information ecosystems that support accessibility, searchability, interoperability, and sustainable document reuse (Sivankalai & Balachandran, 2025; de Oliveira et al., 2023).

However, digitization does not inherently guarantee effective information management. Many scanned documents remain stored as static images or image-based PDFs. Although electronically archived, these files cannot be automatically searched, indexed, or computationally analyzed. Without text recognition capabilities, scanned documents remain passive and fail to support dynamic information retrieval within digital archival systems (de Oliveira et al., 2023; Correia & Luck, 2023). Optical Character Recognition (OCR) technology bridges this gap by converting scanned documents, digital images, or photographs into machine-readable, editable text. OCR enables archival documents to be indexed, keyword-searched, analyzed, and integrated into broader document management systems. In the

context of public sector archives, OCR transforms static scans into active digital assets that can be leveraged for administrative workflows, policy analysis, and public service delivery (Sivankalai & Balachandran, 2025).

Recent advancements have produced a diverse ecosystem of OCR tools, each employing distinct algorithmic architectures and processing pipelines. Tools such as Tesseract OCR, Google Vision OCR, ABBYY FineReader, and Microsoft OCR vary significantly in recognition accuracy, processing speed, layout analysis capabilities, and integration potential. These performance differentials are particularly consequential for government archives, which frequently contain complex structural elements, including varied typography, official seals, handwritten signatures, tabular data, inconsistent scan quality, and heterogeneous administrative formats (Abdallah et al., 2024; Fleischhacker et al., 2025).

Prior research demonstrates that OCR accuracy is highly sensitive to source document quality, scan resolution, noise, page skew, layout complexity, and the efficacy of pre- and post-processing pipelines (Gupta et al., 2007; Drobac & Lindén, 2020). While algorithmic development and single-tool evaluations dominate the literature, systematic comparative studies evaluating multiple OCR systems within real-world local government archival contexts remain scarce. Public institutions require empirical evidence to select tools that optimally balance text extraction accuracy with operational efficiency (Correia & Luck, 2023; Vercruysse et al., 2025).

This study addresses this gap by evaluating OCR performance on administrative documents from government agencies in Malang City. The city was selected due to its high volume of diverse archival outputs, including official correspondence, decrees, service forms, activity reports, and policy documents. While some have undergone scanning, their limited machine-readability restricts archival utility. Malang thus provides a realistic testing environment reflective of public sector digitization challenges.

The primary contributions of this research are twofold: (1) it provides a comparative, metric-driven evaluation of OCR tools using standardized accuracy and efficiency indicators, and (2) it contextualizes findings within the operational realities of local government archives, offering a decision framework for tool selection. Specifically, this study addresses the following research questions: How do selected OCR tools differ in text extraction accuracy, as measured by CER and WER?; How do they differ in processing efficiency, as measured by document processing time?; What tool selection strategy optimally balances accuracy and efficiency for government archival management?.

2. Literature Review

Digital Archives and the Imperative for Machine-Readable Documents

Digital archives constitute electronically stored information assets that can be accessed, managed, and preserved through computational systems. In modern governance, they serve not only as repositories but as instruments of transparency, accountability, and service delivery. Digitization facilitates efficient storage, duplication, distribution, and retrieval relative to physical archives (Sivankalai & Balachandran, 2025). However, true archival utility requires more than image conversion; documents must be searchable, machine-readable, and interoperable with information systems. OCR is therefore a foundational technology for unlocking the analytical and administrative potential of digital archives (de Oliveira et al., 2023; Correia & Luck, 2023). In archival science, document quality is evaluated across dimensions of accessibility, discoverability, interoperability, and long-term usability. Without OCR, scanned documents remain visually accessible but computationally inert. With OCR, they become active data sources capable of supporting keyword retrieval, textual analysis, and evidence-based decision-making.

Optical Character Recognition in Document Processing

OCR is a computational technology that detects and converts textual characters from raster images into encoded, machine-readable text. It functions as a critical interface between image-based documents and text-driven information systems. The OCR pipeline typically comprises four stages: (1) preprocessing (noise reduction, skew correction, contrast enhancement, binarization), (2) segmentation (layout analysis, line/word/character isolation), (3) character recognition (pattern matching or neural classification), and (4) post-processing (lexical correction, language modeling, structural validation) (Nagy, 2016; Drobac & Lindén, 2020). Input quality directly influences output reliability; degraded scans, poor resolution, or complex layouts frequently induce recognition errors (Gupta et al., 2007).

Contemporary OCR systems have evolved from rule-based and template-matching approaches to deep learning architectures capable of contextual understanding, layout preservation, and multimodal recognition. Modern tools increasingly parse tables, forms, stamps, and mixed media capabilities essential for government documents that routinely combine structured and unstructured elements (Abdallah et al., 2024; Fleischhacker et al., 2025).

OCR Accuracy: Character Error Rate and Word Error Rate

OCR accuracy quantifies the fidelity of extracted text relative to a validated reference (ground truth). It is a critical determinant of archival usability, as extraction errors propagate into search indexing, data analytics, and administrative workflows. Accuracy is conventionally measured using Character Error Rate (CER) and Word Error Rate (WER), both derived from edit distance metrics (typically Levenshtein distance) that quantify substitutions (S), deletions (D), and insertions (I) relative to the reference length (N):

$$CER/WER = \frac{S+D+I}{N}, \quad \text{(i)}$$

CER operates at the character level, capturing fine-grained transcription errors, while WER evaluates lexical integrity. Lower values indicate higher accuracy. In public sector contexts, even minor errors in names, dates, reference numbers, or legal terminology can compromise document validity, impede retrieval, and introduce administrative risk (Sangiacomo et al., 2022; de Oliveira et al., 2023; Singh & Middleton, 2025).

OCR Efficiency and the Accuracy–Efficiency Trade-off

Efficiency in OCR evaluation refers to the computational time required to process a document into searchable text. For government agencies managing high-volume archival backlogs, processing speed directly impacts digitization throughput and operational feasibility. Theoretical and empirical literature consistently identifies a trade-off between accuracy and efficiency: complex recognition models (e.g., deep neural networks with layout analysis) typically yield higher accuracy but require longer inference times, whereas lightweight or cloud-optimized models prioritize speed at potential accuracy costs (Correia & Luck, 2023; Fleischhacker et al., 2025). Consequently, OCR evaluation must adopt a multi-criteria framework that simultaneously assesses accuracy, processing time, and contextual suitability.

3. Research Method

This study employs a comparative quantitative approach with an evaluative design. The quantitative paradigm enables objective measurement of OCR performance using numerical indicators (CER, WER, processing time), while the comparative design facilitates direct performance benchmarking across multiple tools under identical document conditions. The evaluative component assesses practical suitability for public sector archival workflows.

Research Objects and Sampling

The study evaluates four OCR tools: Tesseract OCR (open-source), Google Vision OCR (cloud-based), ABBYY FineReader (commercial desktop), and Microsoft OCR (integrated cloud/desktop). The analysis units consist of scanned administrative documents sourced from government agencies in Malang City. Document types include official letters, decrees, service forms, activity reports, and policy memoranda. Documents were selected using purposive sampling based on the following criteria: (1) representativeness of government administrative archives, (2) availability of verifiable ground truth text, (3) variation in scan quality and layout complexity, and (4) compatibility with all tested OCR tools. This sampling strategy ensures ecological validity and aligns with best practices in OCR benchmarking (Martínek et al., 2020; Fleischhacker et al., 2025).

Data Collection Procedure

Data collection followed a standardized five-stage protocol: Document Preparation: Scanned archives were compiled and standardized for testing.Ground Truth Creation: Each document was manually transcribed to establish a verified reference text. OCR Processing: Each document was processed sequentially through all four tools under identical parameter settings. Output Documentation: Extracted text outputs were saved and version-controlled. Time Recording: Processing duration for each tool-document pair was recorded using system-level timestamping. Ground truth validation and time tracking follow established OCR evaluation standards (Drobac & Lindén, 2020; Vercruysse et al., 2025).

Variables and Measurement

The dependent variable is OCR performance, operationalized across two dimensions: Accuracy: Measured via CER and WER. Lower values indicate higher accuracy. Efficiency: Measured via document processing time (seconds). Shorter durations indicate higher efficiency. The independent variable is the OCR tool type (four levels). The combination of accuracy and efficiency metrics ensures comprehensive performance assessment beyond single-dimension evaluation (Sangiacomo et al., 2022; Correia & Luck, 2023).

Data Analysis

Data were analyzed using descriptive statistics (means, standard deviations) and inferential statistics (One-Way ANOVA). ANOVA was selected to test for statistically significant mean differences across the four OCR tools for each performance metric. Assumptions of normality and homogeneity of variance were verified prior to analysis. The significance threshold was set at $\alpha = 0.05$. Statistical analysis was conducted using SPSS v28.

4. Results and Discussion

Descriptive Statistics of OCR Tool Performance

Performance variations across the four OCR tools are summarized in Table 1. Commercial and cloud-based systems generally demonstrated superior recognition capabilities, particularly for documents with complex layouts or structural variations.

Table 1. Descriptive Statistics of OCR Tool Performance.

OCR Tool	Mean CER	Mean WER	Mean Processing Time (s)	SD CER	SD WER	SD Time
Tesseract OCR	0.112	0.184	5.21	0.021	0.033	0.87
Google Vision OCR	0.054	0.091	3.44	0.012	0.018	0.65
ABBYY FineReader	0.041	0.078	4.12	0.010	0.015	0.72
Microsoft OCR	0.063	0.102	3.89	0.014	0.021	0.69

ABBYY FineReader yielded the lowest CER (0.041) and WER (0.078), indicating superior text recognition accuracy. Google Vision OCR achieved the shortest mean processing time (3.44 s), reflecting optimal computational efficiency. Microsoft OCR demonstrated a balanced profile (CER: 0.063; WER: 0.102; Time: 3.89 s), while Tesseract OCR recorded the highest error rates and longest processing duration (CER: 0.112; WER: 0.184; Time: 5.21 s).

ANOVA Test Results

One-Way ANOVA was conducted to determine whether performance differences were statistically significant. Results are presented in Table 2.

Table 2. One-Way ANOVA Results for OCR Tool Performance.

Variable	F-value	p-value	Criteria	Interpretation
CER	8.421	0.001	$p < 0.05$	Significant
WER	6.937	0.003	$p < 0.05$	Significant
Processing Time	5.118	0.007	$p < 0.05$	Significant

All p-values fall below the 0.05 threshold, confirming statistically significant differences among OCR tools across accuracy and efficiency metrics. This indicates that tool selection materially impacts digitized archive quality.

Integration of Theory, Indicators, Results, and Interpretation

Table 3 synthesizes theoretical foundations, empirical indicators, key findings, and their interpretive implications.

Table 3. Integration of Theoretical Framework, Empirical Indicators, and Findings.

Theoretical Aspect	Indicator	Key Finding	Interpretation
OCR accuracy is measurable via error rates	CER, WER	ABBYY: CER=0.041, WER=0.078	ABBYY achieves highest accuracy through minimal transcription errors
Machine-readability enables archival utility	Output quality	Low CER/WER correlates with searchability	Accurate OCR directly enhances digital archive functionality
Efficiency enables scalable digitization	Processing time	Google Vision: 3.44 s	Cloud-optimized models support high-throughput archival conversion
Document complexity affects recognition	Performance delta	Significant ANOVA results across all metrics	Tools vary in layout/structure comprehension under real-world conditions

Accuracy–efficiency trade-off exists	Metric comparison	ABBYY most accurate; Google fastest	Optimal tool selection requires alignment with institutional priorities

All p-values fall below the 0.05 threshold, confirming statistically significant differences among OCR tools across accuracy and efficiency metrics. This indicates that tool selection materially impacts digitized archive quality.

Discussions

OCR Tool Accuracy

ABBYY FineReader's superior accuracy aligns with its advanced layout analysis engine and proprietary document understanding models, which excel at parsing mixed-format government documents (Abdallah et al., 2024; Fleischhacker et al., 2025). In administrative contexts where precision in names, reference numbers, and legal terminology is non-negotiable, high accuracy mitigates retrieval failures and administrative risk. Google Vision OCR ranked second in accuracy, maintaining competitive error rates while prioritizing speed, making it suitable for general archival conversion. Microsoft OCR demonstrated moderate accuracy, offering practical utility for agencies embedded in Microsoft ecosystems. Tesseract OCR's higher error rates reflect the limitations of out-of-the-box open-source configurations; however, its performance can be substantially improved through domain-spec.

OCR Tool Efficiency

Google Vision OCR's processing advantage stems from cloud-based distributed inference and optimized neural pipelines, enabling rapid document conversion. This efficiency is critical for agencies managing archival backlogs or requiring near-real-time document indexing. Microsoft OCR's competitive speed (3.89 s) further supports its viability for integrated enterprise workflows. ABBYY FineReader's slightly longer processing time (4.12 s) reflects the computational overhead of its comprehensive layout analysis and post-correction modules, illustrating the accuracy efficiency trade-off. Tesseract OCR's longer duration (5.21 s) indicates suboptimal efficiency under default settings, though local deployment offers data sovereignty and cost advantages for budget-constrained institutions.

Accuracy–Efficiency Trade-off and Tool Selection Strategy

The results confirm a measurable trade-off: no single tool dominates across all metrics. ABBYY FineReader excels in precision, while Google Vision OCR leads in throughput. For legally binding documents, decrees, or high-stakes administrative records, accuracy should be prioritized to prevent retrieval failures and compliance risks. Conversely, for mass digitization initiatives targeting general correspondence or historical backlogs, processing efficiency becomes paramount. Tool selection must therefore be multi-criteria, accounting for document typology, institutional IT infrastructure, data security policies, budget constraints, and operational timelines.

Table 4. Contextual Recommendations for OCR Tool Selection.

Theoretical Aspect	Indicator	Key Finding
Legal documents, decrees, high-value records	ABBYY FineReader	Lowest CER/WER; optimal for precision-critical archives
High-volume/mass digitization	Google Vision OCR	Fastest processing; scalable cloud infrastructure
Microsoft ecosystem integration	Microsoft OCR	Balanced performance; seamless compatibility with enterprise platforms
Limited budget / custom development needs	Tesseract OCR	Open-source, modifiable, cost-effective with preprocessing optimization
Complex layouts (tables, forms, mixed media)	ABBYY FineReader	Advanced structural parsing and layout retention
General administrative documents	Google/Microsoft OCR	Adequate accuracy with high throughput for operational workflows

5. Conclusions

This study demonstrates significant performance differences among OCR tools in processing digital archives at government agencies in Malang City. These differences are evident across accuracy metrics (CER and WER) and efficiency metrics (document processing time). Descriptive analysis reveals that ABBYY FineReader achieves the highest accuracy, whereas Google Vision OCR delivers optimal processing efficiency. One-Way ANOVA confirms that these differences are statistically significant ($p < 0.05$) across all measured variables, validating that tool selection materially impacts digitized archive quality. Synthesizing

these findings with the study's objectives, the results confirm that no single OCR system universally outperforms others; rather, optimal tool selection requires a multi-criteria approach that balances accuracy and efficiency according to document typology and institutional workflow. Practically, these findings contribute a structured decision framework for public sector digitization: ABBYY FineReader is recommended for high-stakes legal and administrative records, Google Vision OCR for high-volume archival conversion, Microsoft OCR for integrated enterprise environments, and Tesseract OCR for cost-effective, customizable deployments. This study is limited by its focus on administrative documents from a single municipality and the absence of controlled variations in scan resolution, hardware specifications, and cloud data security protocols. Future research should expand sample diversity, evaluate hybrid OCR pipelines, and assess the impact of standardized preprocessing and post-processing protocols on recognition fidelity. Such investigations will further advance evidence-based tool selection for scalable, secure, and high-quality digital archival systems.

References

Abdallah, A., Eberharter, D., Pfister, Z., & Jatowt, A. (2024). A survey of recent approaches to form understanding in scanned documents. *Artificial Intelligence Review, 57*(342), 1–38. https://doi.org/10.1007/s10462-024-11000-0

Breuel, T. M. (2008). The OCRopus open source OCR system. *Proceedings of SPIE, Document Recognition and Retrieval XV, 6815*, Article 68150F. https://doi.org/10.1117/12.783598

Clanuwat, T., Bober-Irizar, M., Kitamoto, A., Lamb, A., Yamamoto, K., & Ha, D. (2018). Deep learning for classical Japanese literature. *Journal of Imaging, 4*(5), Article 65. https://doi.org/10.3390/jimaging4050065

Correia, S., & Luck, S. (2023). Digitizing historical balance sheet data: A practitioner's guide. *Explorations in Economic History, 87*, Article 101475. https://doi.org/10.1016/j.eeh.2022.101475

de Oliveira, L. L., Vargas, D. S., Alexandre, A. M., Cordeiro, F. C., Gomes, D. D., Rodrigues, M. D., et al. (2023). Evaluating and mitigating the impact of OCR errors on information retrieval. *International Journal on Digital Libraries, 24*(1), 45–62. https://doi.org/10.1007/s00799-023-00345-6

Drobac, S., & Lindén, K. (2020). Optical character recognition with neural networks and post-correction with finite state methods. *International Journal on Document Analysis and Recognition (IJDAR), 23*(4), 279–295. https://doi.org/10.1007/s10032-020-00359-9

Fleischhacker, D., Kern, R., & Göderle, W. (2025). Enhancing OCR in historical documents with complex layouts through machine learning. *International Journal on Digital Libraries, 26*(1). https://doi.org/10.1007/s00799-025-00413-z

Google Cloud. (2024). *Cloud Vision API documentation*. https://cloud.google.com/vision/docs/ocr

Gupta, M. R., Jacobson, N. P., & Garcia, E. K. (2007). OCR binarization and image pre-processing for searching historical documents. *Pattern Recognition, 40*(2), 389–397. https://doi.org/10.1016/j.patcog.2006.04.043

Martínek, J., Lenc, L., & Král, P. (2020). Building an efficient OCR system for historical documents with little training data. *Neural Computing and Applications, 32*(3), 17209–17227. https://doi.org/10.1007/s00521-020-04910-x

Nagy, G. (2016). Disruptive developments in document recognition. *Pattern Recognition Letters, 79*, 106–112. https://doi.org/10.1016/j.patrec.2015.11.024

Sangiacomo, A., Hogenbirk, H. D., Tanasescu, R., Karaisl, A., & White, N. (2022). Reading in the mist: High-quality optical character recognition based on freely available early modern digitized books. *Digital Scholarship in the Humanities, 37*(4), 1197–1209. https://doi.org/10.1093/llc/fqab106

Singh, L. G., & Middleton, S. E. (2025). Tabular context-aware optical character recognition and tabular data reconstruction for historical records. *International Journal on Document Analysis and Recognition (IJDAR), 28*(3), 357–376. https://doi.org/10.1007/s10032-025-00543-9

Sivankalai, S., & Balachandran, S. (2025). Transforming physical archives into searchable digital libraries with optical character recognition. *Preservation, Digital Technology & Culture, 54*(4), 263–284. https://doi.org/10.1515/pdtc-2025-0025

Smith, R. (2007). An overview of the Tesseract OCR engine. *Proceedings of the Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, 629–633. https://doi.org/10.1109/ICDAR.2007.4376991

Smith, R. (2009). Hybrid page layout analysis via tab-stop detection. *Proceedings of the 10th International Conference on Document Analysis and Recognition (ICDAR 2009)*, 241–245. https://doi.org/10.1109/ICDAR.2009.64

Vercruysse, B., Birkholz, J. M., Chandrasekar, K. K., Muheki, D., Thiery, W., Verbeeck, H., et al. (2025). Human-in-the-loop tabular data extraction methods for historical climate data rescue. *International Journal on Document Analysis and Recognition (IJDAR), 29*(1), 1–15. https://doi.org/10.1007/s10032-025-00524-y