JOURNAL OF GREEN COMPLEX ENGINEERING, 2026, Vol.4, No.1, 57~64
https://doi.org/10.59810/greenplexresearch.v4i1.284
e-ISSN 3025-9673

Turning Turnitin Highlights into Structured Data with Report-Aware HSV Segmentation and GPU-Accelerated OCR

Mutiara Yusuf*, Rizki Yusliana Bakti, Titin Wahyuni

Informatics Engineering Program, Faculty of Engineering, Universitas Muhammadiyah Makassar, Makassar, Indonesia

ABSTRACT
Turnitin similarity reports contain rich spatial, color, and textual information, yet highlighted passages are usually inspected visually rather than converted into structured data. This study develops a report-aware pipeline that transforms color-coded Turnitin highlights into analyzable records. The method combines PDF rendering, page-type triage, hue–saturation–value (HSV) segmentation, morphological processing, connected-component analysis, geometric filtering, and GPU-accelerated optical character recognition (OCR), with PDF text-layer inspection as supporting information. The evaluation corpus comprised 106 reports and 3,057 pages. Page triage identified 1,315 highlight-bearing pages and routed 56.98% of pages away from detailed processing. HSV segmentation produced eight color classes and 11,471 pre-filter components. After filtering and text extraction, 7,530 structured records were retained; 76.75% satisfied the predefined good-readability criterion, 20.04% were very short fragments, and 3.21% were noisy, garbled, or empty. Document workload was strongly right-skewed. The results show that visual similarity annotations can be operationalized as auditable document data without treating textual similarity as automatic evidence of plagiarism.

ARTICLE HISTORY
Received June 18, 2026
Received in revised form July 19, 2026
Accepted August 20, 2026
Available online August 27, 2026.

KEYWORDS
Turnitin, HSV color segmentation, optical character recognition, document analysis, similarity analytics

1. Introduction

Academic integrity is increasingly mediated by digital systems, but a similarity score is not a self-interpreting measure of misconduct. Students can generate overlap through direct copying, weak paraphrasing, conventional terminology, quotations, templates, or citation practices, and their understanding of these forms is uneven. Recent evidence shows that students distinguish obvious copying more consistently than less explicit forms of reuse (Wilson et al., 2025), while institutional responses are most defensible when detection is combined with awareness, monitoring, and evaluation (Guruge & Kadel, 2023). AI-supported assessment adds another layer: lecturers value automated feedback and analytics but remain concerned about privacy, bias, and the erosion of professional judgment (Choiriyah et al., 2025). For computational research, the implication is clear: similarity evidence may be automated, but its interpretation must remain separable from a misconduct decision.

Most plagiarism-related computing studies focus on comparing content. OCR has been used to bring handwritten assignments or image-based submissions into plagiarism workflows (Kodali et al., 2023; Kavatage et al., 2023; Chaudhari et al., 2024), while multimodal proposals increasingly combine OCR and semantic processing to accommodate scans, handwriting, text, and AI-generated content (Kumar et al., 2025; Chandel et al., 2026). These studies address a necessary problem—making diverse student work comparable—but they usually begin with raw documents and attempt to estimate similarity. The present study starts at a different point in the workflow: a commercial similarity report has already been generated, and the objective is to convert its visual evidence into structured data.

A Turnitin report contains more than an overall percentage. Its pages encode the location of matching passages, match-group colors, compact source identifiers, and the textual context surrounding each highlight. In routine use, these elements are inspected manually, one page and one passage at a time. Manual review is suitable for individual decisions, but it does not naturally support cohort-level analytics, reproducible auditing, or construction of paragraph-level datasets. The technical gap is therefore one of document engineering: how can color, geometry, page context, and text be recovered from a multi-page report while preserving provenance?

Several studies establish individual building blocks. Image-borne text inside scientific PDFs can be extracted by OCR so that it does not escape similarity checking (Sharmeen et al., 2022), and image processing can reveal character-injection strategies that interfere with text-based document comparison (Namburu et al., 2022). Generic PDF-analysis systems likewise emphasize that useful extraction depends on structural decomposition rather than a single recognition call (Amutha et al., 2024; Raval & Bhaidasna, 2025; Ospan et al., 2026). These findings motivate a staged architecture in which irrelevant pages are removed early, color regions are localized, non-target graphics are filtered geometrically, and only then is OCR applied.

The closest methodological precedent is highlighted-text extraction itself. Lee et al. reported that HSV segmentation combined with OCR can recover Korean-English highlighted text and evaluated the pipeline using intersection-over-union and character-level recognition metrics (Lee et al., 2026). This evidence establishes that colored highlights are computationally recoverable. The novelty requirement for the present study is therefore narrower and more defensible: rather than claiming a new combination of HSV and OCR, this work extends highlighted-

CONTACT: Mutiara Yusuf 105841114722@student.unismuh.ac.id Informatics Engineering Program, Faculty of Engineering, Universitas Muhammadiyah Makassar, Makassar, Indonesia

text extraction to a report-aware, multi-page parsing problem that includes page triage, eight match-group color classes, source-ID rejection, record provenance, and highly heterogeneous document workloads.

HSV is attractive because Turnitin's markup uses color as a visual code and because hue can be separated from brightness. Nevertheless, color-space selection is task dependent. Comparative studies report different behavior for RGB, YCbCr, XYZ, Lab, and HSV under different segmentation strategies (Abdelsadek et al., 2022; Al Garea et al., 2024; Armağan, 2022), while threshold-based processing remains sensitive to implementation and platform details (Budak et al., 2022). For controlled report rendering, a lightweight and interpretable HSV pipeline is therefore plausible, but its thresholds should not be assumed to generalize automatically across interfaces, export modes, or scanned copies.

Once extracted, structured text can feed later similarity analysis. Contemporary approaches include lexical ranking, embedding-based document comparison, and hybrid semantic models (Narsale et al., 2025; Paladugu & Kancherla, 2024; Zadgaonkar et al., 2024; Bhoyar et al., 2025). Madoa et al. (2026) applied TF-IDF and cosine similarity to retrieve relevant articles, illustrating how structured textual representations can support downstream similarity analysis. In the present study, such lexical or semantic methods are positioned after extraction: the contribution here is the upstream acquisition layer that converts visually highlighted report content into structured, provenance-preserving text records.

The manuscript therefore defines its contribution as auditable data acquisition, not automated plagiarism adjudication. This tighter positioning avoids overstating the meaning of a Turnitin flag and distinguishes the engineering achievement—recovering structured evidence—from the later normative decision about whether a passage is acceptable.

The study addresses four objectives: (1) quantify page-level workload that can be routed away before expensive processing; (2) characterize the HSV color-component distribution; (3) assess the usability of extracted text using a transparent readability taxonomy; and (4) quantify document-level workload heterogeneity relevant to scalable deployment. The working hypothesis is that a report-aware pipeline can transform a large proportion of visual similarity annotations into readable, provenance-preserving records while avoiding unnecessary processing of non-highlight pages.

2. Methods

2.1 Study design and data source

The evaluative corpus comprised 106 Turnitin PDF reports generated for student proposals and undergraduate theses in an Informatics Study Program at an Indonesian university. Together, the reports contained 3,057 pages. The PDFs included native or rendered document text, color-coded similarity highlights, source-ID elements, and report-specific graphics. The document was therefore treated as a multimodal artifact: the text layer carried lexical content, while the rendered page carried spatial and color information that plain-text extraction would not preserv.

Analytical units differed across stages. The document-level denominator was 106 reports; page classification used 3,057 pages; HSV processing generated 11,471 pre-filter color components; and post-filter extraction yielded 7,530 structured highlight records. These denominators were kept separate because a pre-filter connected component is not equivalent to a final text record, and each page can contain multiple components and multiple records. The source dataset available for this manuscript provided aggregate outputs rather than manually annotated masks or line-by-line transcriptions. Consequently, no unreported precision, recall, F1-score, IoU, character error rate (CER), or word error rate (WER) values were inferred

2.2 Pipeline architecture

Figure 1 summarizes the workflow. Each PDF page was rendered at approximately 200 dpi to establish a consistent image representation for color analysis while preserving the original PDF for text-layer inspection. A page-type triage stage first identified pages likely to contain relevant highlights. Candidate pages were converted from RGB to HSV and pixels were assigned to one of eight empirically calibrated classes: red, orange, yellow, green, cyan, blue, purple, and pink.

Binary color masks were processed with morphological closing to reconnect fragments belonging to the same highlighted line. Connected-component analysis then converted masks into candidate regions. This staged design is consistent with lightweight computer-vision systems that combine HSV segmentation, connected components, and geometric filtering (Bio & Santos, 2026) and with document-recognition pipelines that separate localization from transcription (Ospan et al., 2026; Amutha et al., 2024). Candidate regions were filtered by size and aspect ratio to suppress compact source-ID labels and other colored artifacts before OCR

EasyOCR with CUDA-enabled GPU execution was used as the principal recognition engine, and the PDF text layer was inspected with pdfplumber as supporting information about text availability. GPU-enabled recognition is appropriate when one page can generate many text crops and when a small number of documents create large recognition batches; multi-stage computer-vision systems similarly use GPU OCR as a distinct downstream module (Bhatt et al., 2026). Final records were designed to preserve document identifier, page number, color class, bounding-box coordinates, extracted text, and an operational text-quality category.

2.3 Page-type triage

Page triage used the fraction of pixels satisfying calibrated color criteria. If N_c denotes colored pixels and $H \times W$ is page area, the colored-pixel ratio was $r_{color} = N_c/(H \times W)$. Pages with $r_{color} < 0.0005$ (0.05%) were classified as no_highlight; pages from 0.0005 to 0.005 were classified as low_highlight; and pages above 0.005 (0.5%) were classified as content, subject to a separate cover_or_image rule for large graphical objects. Content and low_highlight pages were retained as highlight-bearing classes.

The purpose of triage was routing, not semantic classification. Early exclusion of irrelevant content is a common scalability principle in document analysis. Paragraph filtering can substantially reduce long documents before similarity modeling (Makawana & Mehta, 2023), and specialized scholarly-document extraction similarly benefits from identifying relevant structures before applying expensive downstream models (Raval & Bhaidasna, 2025).

2.4 HSV segmentation and region filtering

RGB values were normalized to [0,1]. For each pixel, $C_{max} = \max(R', G', B')$, $C_{min} = \min(R', G', B')$, and $\Delta = C_{max} - C_{min}$. Value was $V = C_{max}$; saturation was zero when $C_{max} = 0$ and otherwise Δ / C_{max}; hue was calculated piecewise according to the maximum channel. The implementation used the OpenCV convention in which hue is represented on a 0–179 scale and saturation/value on 0–255 scales.

Thresholds were calibrated empirically from report highlights. This approach is computationally efficient and transparent under a controlled

visual grammar, but comparative segmentation research shows that color representation and preprocessing alter region quality (Abdelsadek et al., 2022; Al Garea et al., 2024). Adaptive segmentation can improve robustness when appearance varies (AlRifaee et al., 2024), while hybrid color and deep-learning systems offer stronger adaptability at additional computational cost (Ananthagoma et al., 2025; Andrade et al., 2026). The present study deliberately retained the simpler threshold approach because report colors and rendering resolution were comparatively standardized.

After segmentation, morphological closing used an approximately 15×3 pixel horizontal kernel. Connected components returned bounding boxes (x, y, w, h), and aspect ratio was defined as AR = w/h. Highlight bodies were expected to satisfy approximately $w \geq 80$ pixels, $8 \leq h \leq 60$ pixels, and $AR \geq 4$. Compact source-ID labels were characterized by approximately $w \leq 40$ pixels, $h \leq 40$ pixels, and $AR \leq 2.5$. Morphology and geometry provide interpretable post-segmentation control and are common companions to HSV localization (Chavan et al., 2022; Fataniya et al., 2025). Overlap remains a potential failure mode; overlap-aware segmentation research shows that deduplication becomes necessary when regions intersect (Deng et al., 2024).

2.5 Text extraction, quality taxonomy, and statistics

The retained crops were recognized with EasyOCR. Extraction quality was summarized using five operational categories. Good records had an alphabetic-character ratio of at least 60%; noisy records had a ratio from 40% to 60%; garbled records fell below 40%; tiny records contained fewer than 10 characters; and empty records contained no text. With N_alpha representing alphabetic characters and N_total the total recognized characters, the ratio was r_alpha = N_alpha/N_total. Tiny was kept separate from recognition noise because a correctly read short phrase may still be semantically inadequate for paragraph-level analysis

This taxonomy measured readability, not OCR accuracy. Recognition accuracy requires aligned ground-truth text and metrics such as CER or WER. In other document tasks, OCR errors measurably constrain downstream matching (Arisa et al., 2025) and language-model post-processing is evaluated through WER rather than a heuristic readability ratio (Rohit et al., 2023). Visual perturbations can also reduce machine extractability even when humans retain legibility (Nguyen et al., 2026). Accordingly, the present study reports the 76.75% good category only as an operational usability indicator.

Page-type and text-quality proportions were calculated from reported counts. Descriptive 95% Wilson intervals were computed for selected proportions. Because pages and records were nested within reports, these intervals were not treated as cluster-adjusted population inference. Color diversity was summarized using normalized Shannon entropy $H/\ln(K)$, where $H = -\Sigma p_i \ln(p_i)$, and concentration using the Herfindahl–Hirschman index, $HHI = \Sigma p_i^2$. Document workload was characterized by minimum, maximum, mean, median, and the share contributed by the ten densest reports.

Reproducibility also depends on recording the configuration of every stage. A deployable implementation should retain rendering resolution, HSV bounds, morphological kernel size, connected-component rules, geometric thresholds, OCR language packs, OCR confidence values when available, and software versions alongside the extracted records. This metadata is especially important because apparently minor preprocessing choices can alter segmentation results; comparative color-space studies show that region quality changes with representation and algorithm (Abdelsadek et al., 2022; Al Garea et al., 2024), while threshold implementations can trade accuracy against computational efficiency (Budak et al., 2022). The present workflow controls one major source of variability by rendering at a fixed resolution, but future releases should store normalized bounding boxes in addition to pixel coordinates. Normalization would allow the same record schema to remain interpretable if pages are re-rendered at a different dpi. A further improvement is stage-wise logging: each candidate region should retain whether it was merged by morphology, excluded by geometry, sent to OCR, or recovered from a native text layer. Such provenance would turn error analysis from a manual visual exercise into a queryable diagnostic process and would make comparisons between classical and adaptive segmentation methods substantially more rigorous.

2.6 Interpretive boundary

The pipeline structures similarity evidence; it does not decide plagiarism. A highlighted passage may represent a legitimate quotation, properly cited overlap, common technical language, a template phrase, or problematic copying. Future supervised models should therefore use human-validated labels rather than treating the existence of a highlight as a positive misconduct label. This boundary is consistent with evidence that plagiarism understanding varies across reuse forms (Wilson et al., 2025) and with institutional approaches that combine technology with human evaluation (Guruge & Kadel, 2023).

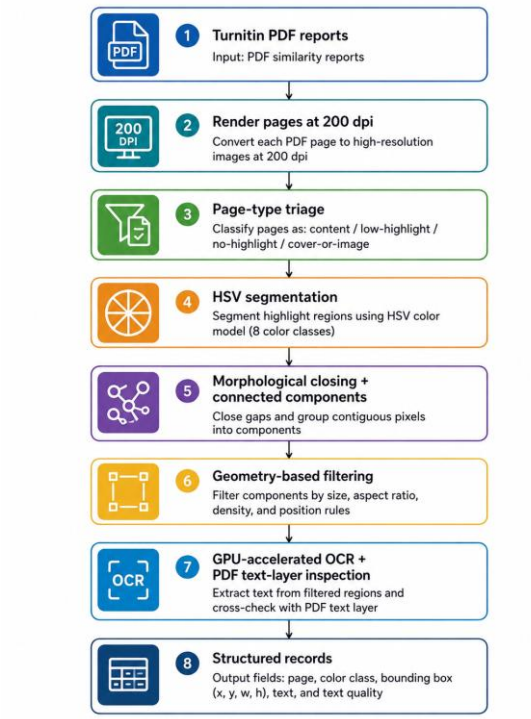

Figure 1. Report-aware workflow for converting Turnitin highlights into structured records.

3. Findings and Discussion

3.1 Page routing and extraction yield

The 3,057 pages were classified as content (637; 20.84%), low_highlight (678; 22.18%), no_highlight (1,737; 56.82%), and cover_or_image (5; 0.16%). Content and low_highlight together represented 1,315 pages (43.02%), whereas 1,742 pages (56.98%) were routed away from detailed extraction. The descriptive Wilson interval for highlight-bearing pages was 41.27–44.78%, and the corresponding interval for routed-away pages was 55.22–58.73%.

This result makes page triage a first-order scalability mechanism rather than a minor preprocessing detail. More than half of the observed page volume did not enter the same segmentation and OCR workload as highlight-bearing pages. The 56.98% fraction should not be translated directly into wall-clock speedup because triage has its own computational cost and later processing cost varies by page. Even so, the structural reduction is substantial and resembles the broader strategy of

removing irrelevant content before expensive similarity modeling (Makawana & Mehta, 2023).

After morphology, connected-component analysis, and geometric filtering, 7,530 final records entered the structured output. The difference between 11,471 pre-filter components and 7,530 final records should not be described as a detection rate: the two stages use different object definitions. The pre-filter count represents connected color regions, whereas the final count reflects regions retained or consolidated after report-specific geometry. This denominator discipline is necessary for reproducible pipeline reporting

3.2 HSVcomponent structure

Across eight calibrated classes, blue produced 4,144 pre-filter components (36.13%), purple 2,064 (17.99%), pink 1,403 (12.23%), green 1,283 (11.18%), orange 846 (7.38%), cyan 742 (6.47%), yellow 717 (6.25%), and red 272 (2.37%). Blue and purple together represented 54.12% of all 11,471 components. Normalized Shannon entropy was 0.870 and HHI was 0.204, indicating meaningful diversity with moderate concentration rather than domination by a single class.

The asymmetry is technically important. Calibration errors in blue or purple would affect a large portion of candidate regions, but the other six classes still represent 45.88% of components. A one- or two-color detector would therefore be unsuitable. At the same time, the result should not be interpreted semantically: Turnitin colors identify match groups, not severity or intent. The segmentation literature similarly cautions that color-space performance depends on task and implementation (Armağan, 2022; Budak et al., 2022), so robustness should be tested under palette and rendering changes rather than inferred from this single environment.

3.3 Text usability and recognition limitations

Of the 7,530 final records, 5,779 were classified as good (76.75%; descriptive 95% Wilson interval, 75.78–77.69%). Tiny fragments accounted for 1,509 records (20.04%; 19.15–20.96%), noisy records for 63 (0.84%), garbled records for 42 (0.56%), and empty records for 137 (1.82%). Noisy, garbled, and empty categories together comprised 242 records, or 3.21% (2.84–3.64%).

The dominant limitation is therefore not simply unreadable OCR. One fifth of final records are short enough to be context poor even when recognized correctly. For any later semantic analysis, these fragments should be mapped back to complete sentences or paragraphs in the native PDF text layer. That step would preserve the visual provenance of the highlight while restoring linguistic context. OCR systems used in assessment and form-processing research show that recognition errors propagate into later scoring or matching (Baclayon et al., 2024; Arisa et al., 2025), so future work should separately quantify localization error, transcription error, and context-reconstruction error.

Formal benchmarking remains necessary. The direct highlighted-text study by Lee et al. used localization and character-level recognition metrics (Lee et al., 2026), whereas the present results only support a readability taxonomy. CER and WER should be calculated on manually transcribed highlight crops stratified by color, crop length, page type, and document quality. Results should also be reported separately for native-text PDFs and image-heavy reports because OCR extractability depends on visual conditions (Nair & Sreekumar, 2024; Nguyen et al., 2026).

The text-quality profile also reveals an error-propagation problem that becomes important once the extraction layer feeds a similarity model. A degraded crop can affect the pipeline in at least three ways: character substitutions alter lexical features, missing words reduce overlap scores, and segmentation errors can detach a phrase from the paragraph that supplies its semantic context. The impact is therefore not linear. A small OCR error in a common word may have little effect, whereas an error in a distinctive technical term can change retrieval or clustering behavior. Arisa et al. showed that OCR quality directly constrained subsequent fuzzy matching (Arisa et al., 2025), and Rohit et al. demonstrated that post-recognition language modeling can reduce word-level error (Rohit et al., 2023). For this reason, a future Turnitin parser should not expose only a cleaned text string. It should preserve the raw OCR output, any corrected form, the crop image, and a confidence or validation status. Downstream similarity experiments can then quantify how much performance changes when low-confidence regions are removed, corrected, or mapped back to native PDF text. That design would separate recognition uncertainty from the performance of the similarity algorithm itself.

3.4 Workload heterogeneity and GPU rationale

Highlight counts were strongly heterogeneous across 106 reports. The minimum was 4 records and the maximum 633; the mean was 71.04 and the median 44.5, producing a mean-to-median ratio of 1.596. The ten densest reports contained 633, 398, 354, 338, 332, 199, 190, 185, 183, and 174 records. Together, they generated 2,986 records, or 39.65% of the final corpus.

This distribution provides a stronger rationale for GPU OCR than an average workload alone. Recognition demand arrives in bursts: most reports are moderate, but a small set can create hundreds of crops. GPU batching is well matched to this queue structure, and related multi-stage vision systems use GPU-optimized OCR as a dedicated downstream component (Bhatt et al., 2026). However, the present manuscript does not claim a measured speed advantage because a controlled CPU-versus-GPU benchmark was not available. Future experiments should report hardware, batch size, rendering time, segmentation time, OCR time, and high-percentile latency.

3.5 From extracted records to similarity analytics

The structured dataset creates an interface between document vision and later similarity analysis. Each record can retain text, page, bounding box, and match-group color. That provenance enables aggregation by sentence, paragraph, document, or cohort while preserving a route back to the visual evidence. Semantic comparison methods can then be applied after acquisition rather than being entangled with the extraction stage. Current approaches range from TF-IDF-based plagiarism detection (Narsale et al., 2025) to embedding-based similarity (Paladugu & Kancherla, 2024; Zadgaonkar et al., 2024) and contextual PDF comparison (Bhoyar et al., 2025). Guo et al. further distinguish copy-paste from rewriting plagiarism, showing why a single lexical score may not capture all reuse patterns (Guo et al., 2025).

Madoa et al. (2026) used TF-IDF and cosine similarity for article retrieval. An analogous lexical or semantic model could operate on the present dataset only after highlighted regions have been converted into clean, contextualized text. This distinction keeps the extraction problem separate from later similarity scoring: HSV segmentation and OCR are acquisition stages, whereas lexical or semantic models would constitute a subsequent analytical layer.

A mature system should expose both the extracted evidence and its uncertainty. A downstream interface should allow users to inspect the original region, recovered text, surrounding paragraph, and any similarity score rather than presenting a binary plagiarism label. This design principle preserves auditability and reduces the risk that an engineering convenience becomes an unsupported disciplinary judgment.

Generalizability should be examined at the level of the report interface rather than assumed from the number of documents. The 106-report corpus is heterogeneous in highlight density, but the reports originate from one institutional context and appear to share a common Turnitin rendering grammar. A new interface version could alter saturation, source-label geometry, typography, or the position of ancillary graphics without changing the underlying similarity semantics. Static thresholds would then face domain shift even though the conceptual task remained identical. Adaptive color localization (AlRifaee et al., 2024) and hybrid segmentation (Ananthagoma et al., 2025) are plausible responses, but a more economical strategy may be automatic calibration from a small set of known interface colors followed

by the same interpretable component pipeline. The appropriate validation design is therefore cross-template testing: train or calibrate on one report style, test on another, and report both localization and transcription degradation. Such experiments would determine whether the current architecture is genuinely report-aware or merely tuned to a particular export appearance.

3.6 Comparison with related vision approaches and limitations

The current architecture favors classical computer vision because the report uses a relatively standardized visual grammar. Bio and Santos combine HSV multi-color segmentation, connected components, and geometric filtering in a lightweight real-time setting (Bio & Santos, 2026), while Chavan et al. and Fataniya et al. use morphology to stabilize HSV-derived masks (Chavan et al., 2022; Fataniya et al., 2025). These precedents support the engineering logic of a transparent region-processing cascade. Learned segmentation can outperform classical thresholding under more complex scenes, but with higher data and computational requirements (Andrade et al., 2026; Ananthagoma et al., 2025). For a report parser, the appropriate criterion is therefore not maximum model complexity but the simplest method that remains robust across realistic report variation.

Five limitations are central. First, no manually annotated mask benchmark is available, so IoU, precision, recall, and F1 cannot be reported. Second, the readability taxonomy is not equivalent to CER/WER. Third, static HSV thresholds may shift with interface updates, compression, scanning, or altered color management. Fourth, pixel-based geometry depends on the fixed 200-dpi rendering scale. Fifth, overlapping highlight colors can generate ambiguous components, a known segmentation problem for which explicit intersection handling may be required (Deng et al., 2024). These limitations constrain claims to extraction yield, composition, and operational usability.

Future validation should therefore use a stratified ground-truth sample. Each sampled page should be manually annotated for highlight masks, color class, source-ID objects, and exact text. Classical HSV thresholds should be compared with adaptive segmentation (AlRifaee et al., 2024) and, where justified, hybrid color/deep-learning approaches (Ananthagoma et al., 2025). OCR evaluation should use CER and WER, and document-level resampling should account for clustering. Only after this validation should paragraph-level similarity models or predictive decision support be evaluated.

Accordingly, the present evaluation should be interpreted as a systems-level feasibility study rather than a benchmark of detection accuracy. Its strongest evidence concerns throughput reduction, output composition, provenance preservation, and operational text usability. Claims about localization accuracy, transcription correctness, or deployment reliability remain explicitly reserved for a future ground-truth study. This separation reduces the risk of overstating performance while making the next validation step clear and reproducible.

Table 1. Dataset and pipeline accounting.

Stage / unit	n	Share / summary	Interpretation
PDF reports	106	—	Document-level corpus
All pages	3,057	100%	Page-classification denominator
Content	637	20.84%	Substantial highlighted content
Low-highlight	678	22.18%	Sparse but relevant highlights
No-highlight	1,737	56.82%	Routed away from detailed extraction
Cover/image	5	0.16%	Non-target pages
Pre-filter HSV components	11,471	8 color classes	Intermediate segmentation objects
Final highlight records	7,530	100% of final set	Post-filter structured records
Highlights/report	106 reports	Mean 71.04; median 44.5; range 4–633	Right-skewed workload

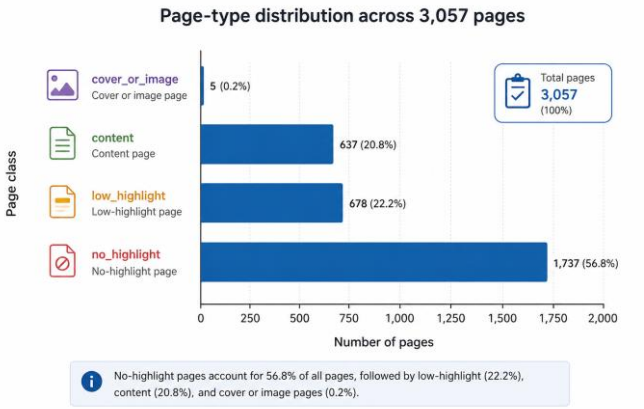

Figure 2. Page-type distribution across 3,057 pages.

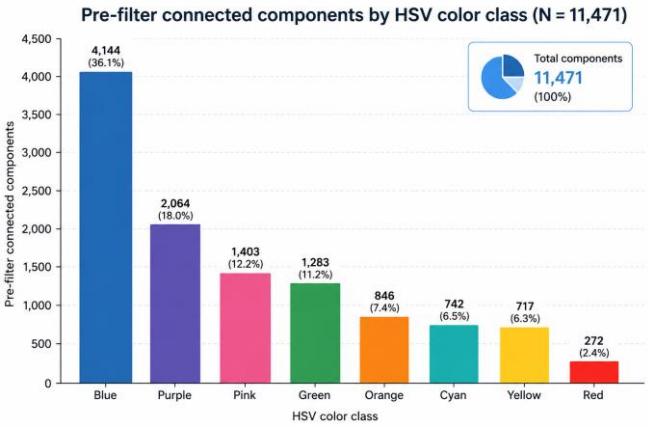

Figure 3. Distribution of 11,471 pre-filter HSV color components.

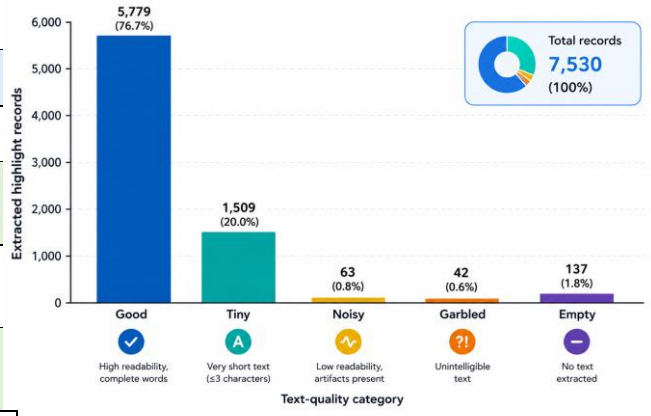

Figure 4. Text-quality profile of 7,530 final structured highlight records.

Table 2. Derived quantitative diagsnostics from reported aggregate counts

Diagnostic	Estimate	95% descriptive interval	Meaning
Highlight-bearing pages	43.02%	41.27–44.78%	content + low_highlight
Pages routed away	56.98%	55.22–58.73%	no_highlight + cover_or_image
Good text records	76.75%	75.78–77.69%	Alphabetic ratio ≥60%
Tiny text records	20.04%	19.15–20.96%	Fewer than 10 characters
Noisy + garbled + empty	3.21%	2.84–3.64%	Degraded or absent text
Blue + purple share	54.12%	—	Largest two HSV classes
Normalized Shannon entropy	0.870	—	Color diversity; 1 = uniform
HSV HHI	0.204	—	Moderate component concentration
Mean/median count	1.596	—	Right-skew indicator
Top-10 concentration	39.65%	—	Share of records in 10 densest reports

From a practical engineering perspective, the structured representation also supports quality-control functions that are difficult with a single similarity percentage. Institutions could audit the distribution of match fragments by document section, identify recurrent short phrases that inflate similarity without substantive concern, or prioritize unusually dense reports for human review. None of these uses requires the system to infer misconduct. They require reliable acquisition, context reconstruction, and transparent linkage to the original report. This distinction is consistent with the broader move from simple lexical matching toward multimodal and semantically informed plagiarism analysis (Kumar et al., 2025; Guo et al., 2025). It also clarifies why the pipeline should be evaluated as document infrastructure: its success is measured by whether it preserves evidence accurately enough for later analysis, not by whether it replaces academic judgment.

4. Conclusion

This study demonstrates that color-coded Turnitin similarity reports can be transformed into structured, machine-readable data through a report-aware pipeline that combines page triage, HSV segmentation, morphology, connected-component analysis, geometric filtering, and GPU-accelerated OCR. Across 106 reports and 3,057 pages, 56.98% of pages were routed away from detailed processing. HSV segmentation produced 11,471 pre-filter components across eight color classes, and post-filter extraction produced 7,530 structured records. Of these, 76.75% satisfied the operational good-readability criterion, while 20.04% were very short and 3.21% were noisy, garbled, or empty.

The principal contribution is an auditable acquisition layer rather than a new plagiarism verdict. The system preserves page, color, location, and text provenance so that later lexical or semantic models can operate on structured evidence. The resulting records can subsequently support lexical or semantic similarity models (Madoa et al., 2026), but the present work remains focused on making visual report evidence computationally usable. Ground-truth localization metrics, CER/WER, cross-version robustness, and human-validated semantic labels are required before the pipeline is used for high-stakes decisions.

References

[1] Abdelsadek, D. A., Al-Berry, M. N., Ebied, H. M., & Hassaan, M. (2022). Impact of using different color spaces on image segmentation. Lecture Notes on Data Engineering and Communications Technologies, 113, 456–471. https://doi.org/10.1007/978-3-031-03918-8_39

[2] Kementerian PPN/Bappenas. Proyeksi kebutuhan pembiayaan revitalisasi sarana dan prasarana pendidikan 2025–2029. Kementerian PPN/Bappenas; 2025.

[3] Peraturan Presiden Republik Indonesia Nomor 12 Tahun 2021 tentang Perubahan atas Peraturan Presiden Nomor 16 Tahun 2018 tentang Pengadaan Barang/Jasa Pemerintah, (2021).

[4] Alanazi AH, Al-Gahtani KS, Alsugair AM, Bin Mahmoud AA, Alsanabani NM. Assessment and ranking of criteria for engineering firm performance using RII, entropy weight method, and TOPSIS. Applied Sciences. 2026;16(11).

[5] Elsherbeny HA, Gunduz M, Ugur LO. A hybrid model for enhancing risk management and operational performance of AEC consultants: An integrated PLS–ANN approach. Sustainability. 2025;17(4):1467.

[6] Ogbu CP, Imafidon MO. Influence of selection criteria on clients' satisfaction with construction consultancy services in Nigeria. Journal of Engineering, Design and Technology. 2022;20(6):1519-37.

[7] Briliyanti A, Rojewski J, Colbry DJL, Colbry KL. Training the trainers: Preparing facilitators to provide professional development for engineers and scientists. ASEE Annual Conference and Exposition Conference Proceedings. 2022.

[8] Bekele AA, Mahesh G. Impact of contractor development programs on the competency of small and medium contractors in Ethiopia. International Journal of Construction Education and Research. 2025;21(1):94-116.

[9] da Alves TCL. The Silo Game: A simulation on interdisciplinary collaboration. Proceedings of the International Group for Lean Construction. 2022. p. 1052–1063.

[10] Ibrahim FSB, Ebekozien A, Khan PAM, Aigbedion M, Ogbaini IF, Amadi GC. Appraising fourth industrial revolution technologies' role in the construction sector: How prepared are construction consultants? Facilities. 2022;40(7–8):515-32.

[11] Kaboré SS, Ngangue P, Soubeiga D, Barro A, Pilabré AH, Bationo N, et al. Barriers and facilitators for the sustainability of digital health interventions in low- and middle-income countries: A systematic review. Frontiers in Digital Health. 2022;4:1014375.

[12] Chung H, Cha K, Lee HS. What drives and hinders the use of new e-customs systems in developing countries of Sub-Saharan Africa? An empirical study from Cameroon. Information and Media. 2023;96:40-64.

[13] Amin B, Ashad H, Maricar H. Kajian peranan konsultan manajemen konstruksi pada proyek konstruksi di Makassar. Jurnal Teknik Industri Terintegrasi. 2024;7(3):1598-605.

[14] Putra IKAA, Pagehgiri J, Ariyanta IPG. Analisis kinerja konsultan pengawas konstruksi dalam pelaksanaan proyek gedung Puskesmas di Kabupaten Tabanan. Jurnal Teknik Gradien. 2021;13(1):48-60.

[15] Taqiudin M, Anggara D, Nukman N. Analisis pengawasan konstruksi: Kajian kinerja konsultan pengawas di proyek gedung RSUD Awet Muda Narmada. Jurnal Ilmiah Telsinas Elektro, Sipil dan Teknik Informasi. 2023;6(2):189-95.

[16] Damayanti T, Yuni NKSE, Yuliana NPI, Wahyudi IGB. Identifikasi kesenjangan kinerja konsultan pengawas konstruksi di Kota Denpasar. Menara: Jurnal Teknik Sipil. 2025;20(1):84-92.

[17] Wulandari A, Dewantoro, Puspasari VH. Analisis kinerja konsultan pengawas konstruksi dalam pelaksanaan proyek di Kota Palangka Raya. Portal: Jurnal Teknik Sipil. 2024;16(2):142-51.

[18] Wijaya A, Andi A, Rahardjo J. Tingkat kepuasan kontraktor terhadap kinerja konsultan manajemen konstruksi di Surabaya. Dimensi Utama Teknik Sipil. 2023;10(2):137-55.

[19] Ali SHSHA, Habtoor AAA, Alnuaimi AS, Šaljić E, Tomašević V, Raut J. A Delphi and Importance–Performance Analysis framework for fire safety competencies of architects and fire safety engineering consultants in the UAE. Buildings. 2026;16(12).

[20] Abas D, Ahadian ER, Saputra MT. Analisis kepuasan pengguna jasa terhadap kinerja konsultan pengawas pada pekerjaan konstruksi di Kota Ternate. Journal of Science and Engineering. 2021;4(2):120-31.
[21] Ogourtsova T, Gonzalez M, Zerbo A, Gavin F, Shikako K, Weiss J, Majnemer A. Lessons learned in measuring patient engagement in a Canada-wide childhood disability network. Research Involvement and Engagement. 2024;10(1).
[22] Duque Monsalve LF, Navarrete Valladares CP, Sandoval Díaz J. Relationship between political participation and community resilience in the disaster risk process: A systematic review. International Journal of Disaster Risk Reduction. 2024;111:104751.
[23] Nooteboom LA, Mulder EA, Vermeiren RRJM, Eilander J, Van Den Driesschen SI, Kuiper CHZ. Practical recommendations for youth care professionals to improve evaluation and reflection during multidisciplinary team discussions: An action research project. International Journal of Integrated Care. 2022;22(1).
[24] Nabelsi V, Plouffe V, Leclerc MC. Barriers to and facilitators of implementing overnight nursing teleconsultation in small, rural long-term care facilities: Qualitative interview study. JMIR Aging. 2025;8:e71950.
[25] Chigara B, Moyo T. Factors affecting the delivery of optimum health and safety on construction projects during the COVID-19 pandemic in Zimbabwe. Journal of Engineering, Design and Technology. 2022;20(1):24-46.
[26] Hong S, Chae GY, Lim HS. Revitalizing regional industries through public innovation intermediaries: A multi-stage evaluation of the maritime Open Lab model. Sustainability. 2026;18(10).
[27] Marengo LL, Soares AN, Romão HRDS, de Araújo DLA, Zilber SN. The evolution of the local innovation agents program methodology and its contribution to innovation management. REGEPE Entrepreneurship and Small Business Journal. 2022;11(2).

This page is intentionally left blank.

Journal of Green Complex Engineering