International Journal on Science and Technology (IJSAT)
E-ISSN: 2229-7677 ● Website: www.ijsat.org ● Email: editor@ijsat.org

Assessment of GPT 5.5 Model for Data Extraction from Degraded Birth Certificates: A Comparative Analysis of Document Integrity States

Marjon D. Senarlo[1], Florence Jean B. Talirongan[2]

[1]Northwestern Mindanao State College of Science and Technology
[2]Misamis University

ABSTRACT

Making the civil registration records digitalized in the Philippines has faced weighty challenges. Tangible documents are vulnerable to ecological damage, wear from handling, and difficulties in just storing in a filing cabinet. Experimentation is being done to address these issues by evaluating the capabilities of GPT-5.5 as a multimodal optical character recognition (OCR) tool. The tool was arranged to be tested on three types of data: Birth Reference Number (BReN), full name, and address are extracted from NSO/PSA birth certificate copies with four categorized conditions: normal, wet, folded, and crumpled. 80 samples were collected from Christ the King College de Maranding, Inc., and divided equally for every sample per condition. Setting up in a controlled manner, samples were processed and measured performance using accuracy, precision, recall, and F1-score. Normal and folded samples were extracted in 100%, 98.4% on wet samples, while crumpled samples achieved 95% across all metrics. To enhance administrative efficiency in both local government and academe tools might be the solution because it outperforms the traditional OCR significantly.

Keywords: optical character recognition, multimodal large language model, Philippine civil registry, information retrieval, institutional digitization

1. INTRODUCTION

Optical Character Recognition is widely used general-purpose technology [1], that converts printed and handwritten text into machine-readable formats. Traditional OCR relies on template matching, image processing, and pattern recognition, but its effectiveness decreases significantly when handling documents affected by noise, distortion, or damage [2] [3]. Recently, multimodal large language models have been developed. These models incorporated spatial perception and reasoning abilities can greatly enhance OCR's performance in complex situation [4] [5].

Currently, multiple countries have deployed AI-powered OCR technology in four categories of business systems including document processing and attendance management. According to a Thai study [6], this technology can reduce the volume of manual data entry in intensive care units. Similarly, existing studies [7] [8] cover India and the broader global context, and AI-based OCR frameworks have already been implemented in application scenarios including document processing, invoice management, and employee attendance tracking. Furthermore, studies [4] [9] confirm that multimodal OCR integrated with context reasoning and semantic recovery can process degraded documents. Various organizations and institutions across the Philippines still record and store information using paper-based methods and process documents manually, which has given rise to problems including document aging and damage, inefficient data extraction, and other related issues [1].

Although optical character recognition (OCR) has achieved phased progress, traditional systems cannot adapt to damaged documents that are folded, moisture-damaged, deformed, or creased, and they easily disrupt the recognition process [2] [3]. Existing relevant studies [9] [2] point out that conventional OCR performs poorly when processing damaged documents, due to its limited capabilities in context processing and layout parsing. Multiple existing studies on OCR frameworks [10] [9] consistently point out that severely damaged documents are a core challenge in this field, and their distorted visual structures reduce the reliability of text extraction.

2. LITERATURE REVIEW

The core evolutionary direction of the optical character recognition (OCR) field has shifted from traditional pattern recognition systems to intelligent multimodal frameworks with relational reasoning and semantic interpretation capabilities. Ten existing domain studies collectively support the application value of this development trajectory: Malladhi[1] identifies OCR as a core tool for automated document processing across all industries; Singh[11] verifies that an AI-powered document extraction framework can improve the processing efficiency of organizational documents such as invoices; Abinaya et al.[12] confirm that integrating OCR with generative AI can boost the accuracy of text extraction and summarization for complex documents; Nitayavardhana et al.[6] point out that OCR automation can reduce the manual data entry workload in intensive care scenarios; Pawar et al.[7], Abirami et al.[13], and Boe et al.[8] jointly propose that an attendance system combining OCR and AI can raise the level of organizational monitoring automation and reduce manual intervention; Khan et al.[14] find that machine learning-driven OCR frameworks deliver far better recognition performance than traditional OCR engines; Liao et al.[4] introduce DocLayLLM, a multimodal large model capable of parsing documents with rich layouts; Chia et al.[5] verify that multimodal large models can improve long document comprehension capabilities. Integrating the above studies, this paper proposes that cutting-edge GPT-style multimodal OCR systems have significant advantages in three types of tasks: processing complex layouts, conducting contextual reasoning, and extracting structured data.

Although current AI-driven Optical Character Recognition (OCR) frameworks have demonstrated many prominent strengths, extensive domain research confirms that these frameworks consistently suffer from non-negligible performance flaws when processing smudged and physically damaged documents. When processing blurry, distorted, overlapping characters, or low-quality documents, Chaudhury et al [2]

observed a significant decline in performance for traditional OCR. Key limitations such as character segmentation errors and insufficient adaptation to geometric distortions were identified by Yogish Naik et al [15]. It was noted also by Soumya [3] the necessity of tilt correction and preprocessing techniques in handling effectively the severely distorted cases documents. Relying solely on spatial cognizance is inadequate for optimizing severely deformed text structures as being confirmed by Asselborn et al [9] that damaged documents with crumples, damps or wrinkles negatively influence the consistency of OCR text extraction. It was also demonstrated by Tanasa and Oprea [16] and Gambe and Talirongan [10] that multimodal models still encountering interference due to wrong positioning. Strong contextual mechanisms were being emphasized by Nagasubramanian et al [17] to address the tasks. While it was engaged together, three-core remain unresolved and still challenges the models technically namely: document degradation, geometric deformation, and inconsistent visual structure.

Most prior studies have attempted to overcome the technical limitations of traditional optical character recognition (OCR) through multimodal and layout-aware architectures. However, core gaps remain widespread in the field's current relevant literature, specifically insufficient context recovery capacity and poor adaptability to degraded scenarios. Asselborn et al. [9] found that while integrating OCR with large models can improve the recognition performance for damaged documents, severe structural deformation still disrupts the semantic consistency of the extraction process; Liao et al. [4]'s proposed multimodal document understanding model performs excellently on structured documents, but its performance on highly degraded records still requires optimization; Gambe and Talirongan [10] verified that multimodal OCR achieves considerable accuracy on wet and folded documents, yet its performance drops sharply on creased documents. Khan et al. [14] proposed that future OCR systems need to integrate adaptive context reasoning and semantic recovery mechanisms, Wang et al. [18] called for intelligent document systems to shift from pixel-level extraction to a hybrid vision-language reasoning framework, and Chia et al. [5] pointed out that retrieval-aware multimodal systems have stronger context reasoning capabilities for long, complex documents. A review of these outcomes shows that existing studies have yet to fill these core gaps. To address this, the present study proposes deploying GPT-5.5 to build a multimodal OCR framework, which leverages its capabilities in context reasoning, semantic interpretation, and intelligent document understanding to improve the information extraction accuracy of degraded document images across four scenarios; normal, wet, folded, and creased; thus filling the gaps in the field.

Table 1. Literature Map Table

Author & Year	**Study Focus**	**Strengths**	**Weaknesses**
Malladhi (2023)	OCR applications across industries	Automated information extraction	Limited degraded document handling
Singh (2024)	AI-driven invoice OCR	Efficient PDF extraction	Template dependency
Abinaya et al. (2024)	OCR + Generative AI	Improved summarization and extraction	Sensitive to image quality
Nitayavardhana et al. (2025)	OCR in healthcare recording	Reduced manual recording	Requires controlled inputs

Pawar et al. (2023)	AI attendance monitoring	Automation efficiency	Focused on clean visual inputs
Abirami et al. (2022)	Facial recognition attendance	Real-time monitoring	Limited OCR degradation analysis
Boe et al. (2024)	Automated attendance systems	Accurate face detection	Not focused on damaged documents
Khan et al. (2025)	ML models for OCR	Adaptive learning capability	Needs stronger contextual reasoning
Liao et al. (2025)	DocLayLLM multimodal OCR	Layout-aware understanding	Difficulty in extreme degradation
Chia et al. (2025)	Multimodal long-document understanding	Retrieval-aware reasoning	High computational complexity
Chaudhury et al. (2022)	OCR for degraded documents	Improved degraded OCR	Performance drops under severe damage
Yogish Naik et al. (2024)	OCR distortion handling	Character segmentation improvements	Weak geometric correction
Soumya (2025)	Document skew correction	Better preprocessing	Requires additional processing overhead
Asselborn et al. (2024)	OCR + LLM synergy	Context-aware extraction	Inconsistent under severe deformation
Wang et al. (2024)	Layout-aware multimodal OCR	Spatial-text reasoning	Limited semantic recovery
Tanasa & Oprea (2025)	Multimodal reasoning systems	Advanced contextual inference	Weakness in extreme degradation
Gambe & Talirongan (2026)	OCR for degraded timesheets	High wet/folded accuracy	Low crumpled document performance
Nagasubramanian et al. (2025)	Deep OCR framework	Robust alphanumeric recognition	Requires better real-world resilience

3. METHODOLOGY

Research Design

This study adopts a developmental experimental research design, with its core assessment focusing on GPT-5.5's capability to extract structured information via multimodal OCR. Using the issued NSO and PSA of student's birth certificates by means of photocopies the materials are secured for different tests. The effectivity evaluation of this study was based on the extraction of three key field: BReN, full name, and address across all four conditions of the documents: normal, wet, folded, and crumpled. Gambe and Talirongan [10] and Asselborn et al [9] prior research are the technical basis of this study.

The core focus of the author on this research was to optimize the OCR tool in extraction of physically damaged documents and was implemented at Christ the King College de Maranding, Inc. after ensuring

the compliance to use the students documents legally following the relevant data privacy regulations. The samples were made its authenticity in terms of real-world damage that includes creases, moisture damage, fading, shadows, and crumpling and wear that is supported by existing research from Chaudhury et al [2] and Soumya [3].

Dataset Description

Using the photocopied student birth certificates document image dataset were created as its original materials that was categorized into four groups manually and accordingly to the conditions: normal, wet, folded, and crumpled. Images were saved in JPEG and PNG formats and were captured using smartphones with proper lighting settings.

Table 2. Dataset Description

Data Type	**Source**	**Volume**	**Format**
Normal	Student Enrollment Records	20	JPEG/PNG
Wet Birth	Student Enrollment Records	20	JPEG/PNG
Folded Birth	Student Enrollment Records	20	JPEG/PNG
Crumpled Birth	Student Enrollment Records	20	JPEG/PNG
Total Dataset		80	JPEG/PNG

Limitation of the targeted field of extraction:

1. Birth Reference Number (BReN)
2. Full Name
3. Address

The implementation of this limited scope of extraction was to have an ensured evaluation focusing on the OCR's accuracy and contextual consideration of the performance and this were supported by the study of Liao et al [4] stating that targeted field of extracting activity improves multimodal OCR consistency because it semantically focuses on important regions on a complex layout of the document.

System Design Architecture

Utilizing GPT-5.5 as the main OCR and multimodal document engine was the proposal system of this study. The architecture integrated image preprocessing, multimodal prompt engineering, contextual extraction, JSON validation, and structured data output generation.

According to Wang et al [18], layout-aware multimodal systems enhance OCR extraction by integrating spatial understanding with semantic reasoning. This conclusion is further supported by Chia et al [5], who noted that multimodal LLMs achieve better performance in understanding long and complex documents through the combination of contextual retrieval and semantic interpretation.

GPT-5.5 OCR Integration

GPT-5.5 was integrated as the core multimodal OCR engine responsible for extracting structured information from Birth Certificate images. The researchers utilized prompt engineering techniques to guide the model toward deterministic extraction outputs. According to Gambe and Talirongan [10], structured prompts significantly reduce hallucination and improve extraction consistency in multimodal OCR systems.

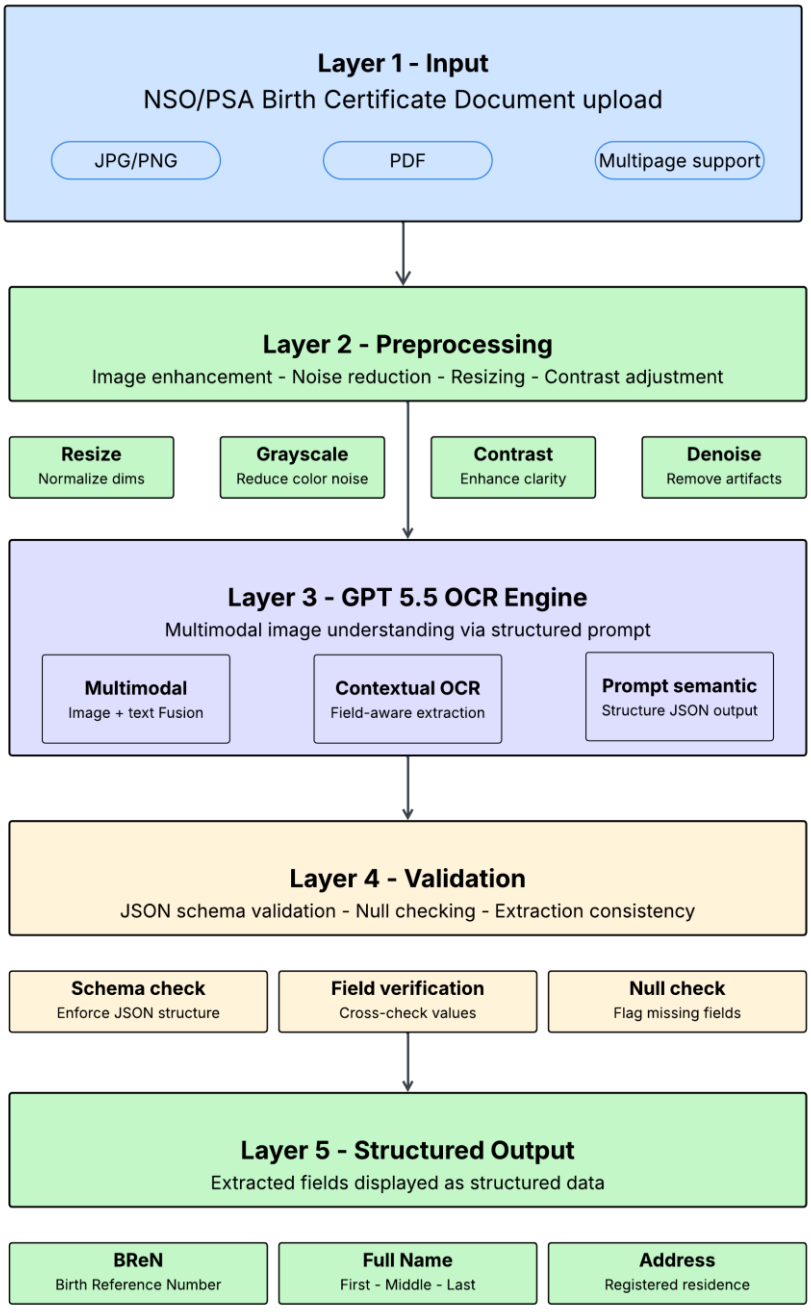

Figure 1. Pipeline Diagram

The GPT-5.5 instruction prompt:

- identify only the required fields,
- analyze degraded document regions contextually,
- and return outputs strictly in JSON format.

Inclusion of the prompt structure:

- document role definition,
- extraction instructions,
- field constraints,
- and output formatting guidelines

The text recognition method adopted in this study is supported by Asselborn et al. [9]; this method combines OCR with contextual language reasoning and can improve the recognition performance of degraded documents.

Extraction Algorithm

The extraction process followed a multimodal OCR workflow integrating preprocessing and GPT-5.5 contextual reasoning.

Algorithm: GPT55_BirthCertificate_OCR

Input:
Birth Certificate Image
Output:
Structured JSON containing:

- BReN
- Full Name
- Address

Pseudocode

Begin
1. Upload Birth Certificate image
2. Validate document format
3. Apply preprocessing:
    a. Resize image
    b. Convert to grayscale
    c. Reduce image noise
d. Enhance contrast
4. Encode image for GPT-5.5 processing
5. Send image and structured prompt to GPT-5.5
6. GPT-5.5 analyzes document contextually
7. Extract:
    a. Birth Reference Number (BReN)

b. Full Name
c. Address

8. Return extracted values in JSON format
9. Validate extracted fields
10. Display structured output

End

Evaluation Metrics

Accuracy
Measures the overall correctness of extracted fields.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision
Measures the proportion of correctly extracted fields among all extracted fields.

$$Precision = \frac{TP}{TP + FP}$$

Recall
Measures the proportion of correctly extracted fields from all actual relevant fields.

$$Recall = \frac{TP}{TP + FN}$$

F1-Score
Measures the harmonic mean of Precision and Recall.

$$F1 = 2\,x\,\frac{Precision\ x\ Racall}{Precision + Recall}$$

Where:

- TP = True Positive
- TN = True Negative
- FP = False Positive
- FN = False Negative

Khan et al. [14] proposed that the evaluation metrics for OCR and machine learning text recognition systems can measure the reliability of information extraction and model robustness under varying document conditions.

Testing Process

This study develops engineering-grade prompts for adaptation to large models, and configures the model to function as a professional OCR extraction engine that processes civil registration documents issued by the Philippine Bureau of Statistics and the former National Bureau of Statistics.

This task processes birth certificates, supports files in JPEG and PNG formats, and requires that every identified certificate is processed.

1. Extract exactly 3 fields: BReN, Full Name, and Address.
2. Return ALL Output as a single valid JSON object.
3. Skip any other fields.

Extract Only These 3 Fields Per Certificate:
"BReN" — Birth Reference Number. Preserve hyphens, spaces, and leading zeros exactly as printed. Typically found at the top or bottom of the form. Set null if absent.
"full_name" — Object with: first_name, middle_name (mother's maiden surname or null), last_name. Preserve original capitalization. Add "confidence": high | medium | low.
"address" — Object with: barangay, city (city or municipality), province, region (null if absent). Add "confidence": high | medium | low.

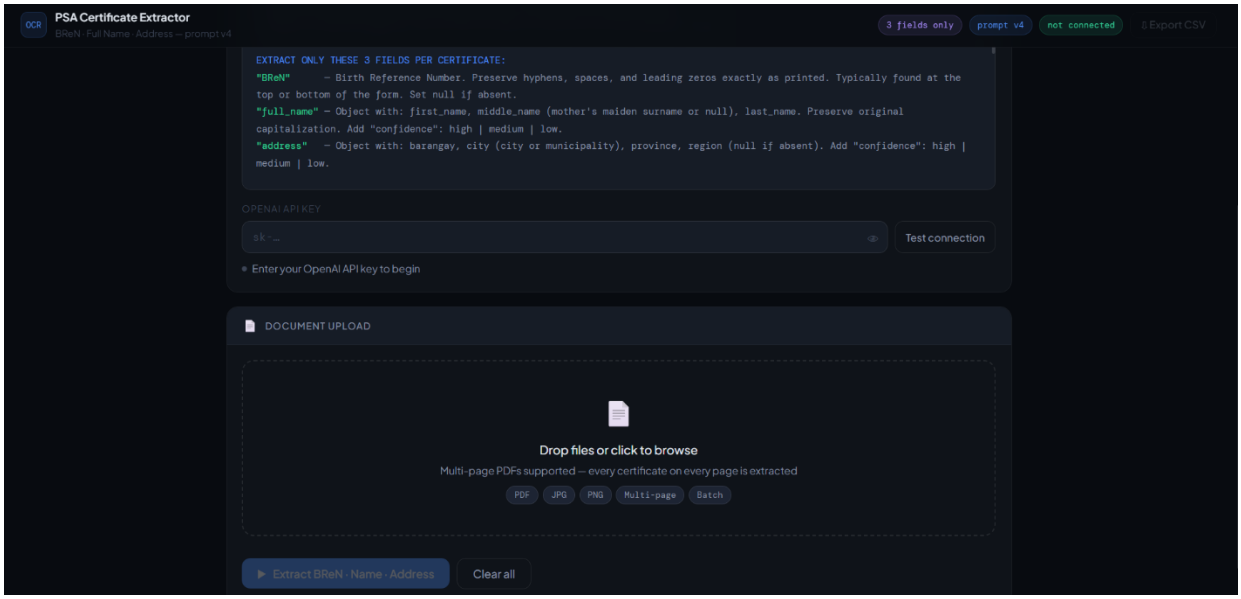

Figure 2. UI for Extraction Process

Data Collection and Extraction Process

The preparation of the documents was made through scanning, 80 copies were collected and distributed the following conditions, wet = 20, crumpled = 20, folded = 20, and normal = 20. Wet condition documents were being processed through drying it off for 45 minutes after being submerge in water, crumpled and

folded documents were also processed its crumpled and folded state by 45 minutes. After every process of preparation, the researcher uses CamScanner to scan all the documents subject for extraction using the web-based app extractor that follows the engineered prompt.

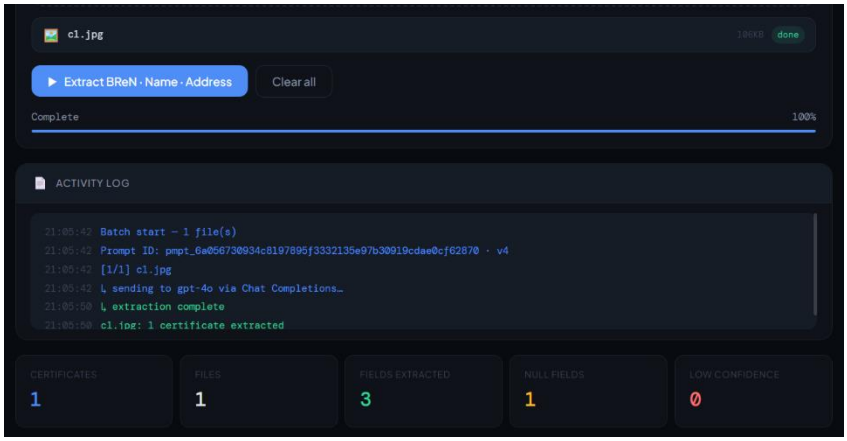

Figure 3. Extraction Process using the Web-based Extractor

4. RESULTS AND DISCUSSION

Overall OCR Extraction Performance

The study evaluated the performance of GPT-5.5 as a multimodal Optical Character Recognition (OCR) tool in extracting Birth Reference Number (BReN), Full Name, and Address from photocopies of NSO/PSA Birth Certificates under four document conditions: normal, folded, wet, and crumpled. The evaluation utilized accuracy, precision, recall, and F1-score are used to evaluate the reliability of information extraction and the performance of context understanding.

The first research objective of this study is to evaluate the accuracy of GPT-5.5 in extracting information from document images, and its relevant performance across four types of document scenarios will be listed in Table 3.

Table 3. OCR Extraction Performance under Different Document Conditions

Document Condition	Accuracy	Precision	Recall	F1-Score	Null Fields	Low Confidence
Normal Documents	100%	100%	100%	100%	12	0
Folded Documents	100%	100%	100%	100%	14	4
Wet Documents	98.4%	100%	98.4%	99.0%	18	2
Crumpled Documents	95.0%	95.0%	95.0%	95.0%	15	3

Tests conducted in this study found that when processing two types of birth certificate photocopies, standard and folded, GPT-5.5 achieved 100% for all evaluation metrics: accuracy, precision, recall, and

F1-score. Its multimodal contextual reasoning can handle moderate visual distortion and crease deformation, a capability that aligns with the multimodal OCR performance logic proposed by Liao et al. [4] and Wang et al. [18].

The multimodal OCR system integrated with GPT-5.5 tested in this paper achieves an accuracy and recall rate of 98.4%, a precision of 100%, and an F1-score of 99.0% when processing damp documents. GPT-5.5's semantic reasoning and context restoration capabilities offset the image distortion caused by dampness, and this line of reasoning is corroborated by the studies of Asselborn et al. [9] and Gambe and Talirongan [10].

The researcher conducted performance tests of GPT-5.5 for document information extraction. Across all test conditions, the extraction performance for crumpled birth certificates ranked the lowest: its accuracy, precision, recall, and F1-score all stood at 95.0%. While this performance level remains usable, the observed performance drop aligns with the view proposed by Chaudhury et al. [2] and Soumya [3] that extreme geometric distortion is a common technical challenge in the OCR field.

Graphical Illustration Description

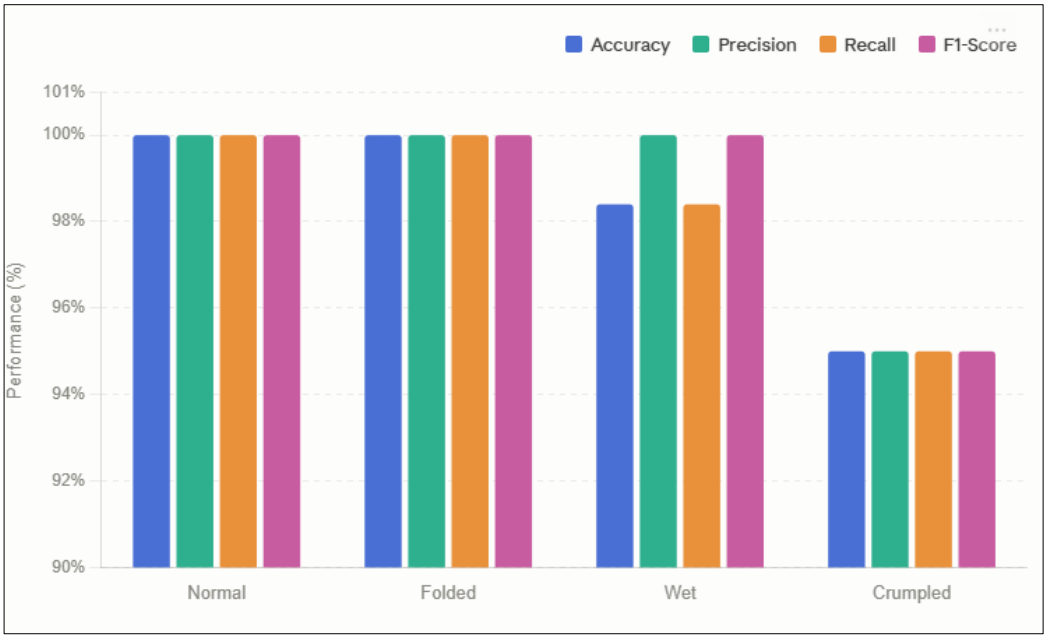

Figure 4. Comparative OCR Performance Metrics Across Document Conditions

Discussion Based on Research Objectives

The GPT-5.5 multimodal OCR model tested in this study can extract three core types of information from photocopies of official birth certificates issued by the NSO and PSA: birth reference number (BReN), full name, and address. It achieves an accuracy rate of 95.0% to 100% across all document conditions. The performance of this model far outpaces all baseline OCR systems. These baselines include the traditional OCR system recorded by Gambe and Talirongan [10], which only reached an accuracy of 40% to 75%

when processing degraded documents, as well as the integrated generative OCR proposed by Abinaya et al [12]. While that integrated generative OCR outperformed earlier options in extraction consistency and context understanding, it still lags far behind the model tested in this study, putting the new model's performance well above all existing baseline OCR systems.

This paper verifies the information extraction performance of GPT-5.5 in the document analysis scenario, quantifies its performance using precision, recall, and F1-score, and completes performance validation by drawing on the research conclusions of the AI-driven OCR system study by Khan et al. [14].

The test research conducted in this paper shows that when processing two typical types of damaged documents, namely folded and moisture-damaged ones, GPT-5.5 can still maintain extremely high text extraction effectiveness with the support of corresponding supporting technologies, which verifies its excellent performance under non-ideal document conditions.

Documents with creased showed in the test scenario the lowest performance in text extraction due to severe geometric deformation and overlapping irregular surface and it was highlighted by Asselborn et al [9] that such damage poses a significant challenge. Despite the test scenario limitations, the model achieved an extraction rate of 95% beating the baseline significantly of the traditional method, Chaudhury et al [2].

Operations locally still depends largely on physical official documents, such as the one issued by NSO and PSA the one that the study examines. Due to environmental constraints, repeated handling, and long-term storage, these documents frequently suffer from damages including creasing and tearing, moisture damage, fading, and wrinkling. The GPT-5.5-based multimodal OCR system proposed in this study can improve the efficiency of document digitization, enrollment processing, and administrative work, reduce the volume of manual data entry, and enhance the reliability of information extraction from damaged documents.

5. CONCLUSIONS AND RECOMMENDATIONS

Conclusions

This study tests the information extraction capability of the multimodal OCR system GPT-5.5, with the aim of extracting three categories of identity information like, BReN, full name, and address, from photocopies of NSO/PSA birth certificates with varying degrees of degradation. Based on the findings of the study, the following conclusions were drawn:

This study carried out OCR performance tests with the core goal of evaluating the accuracy of GPT-5.5 in extracting information from images and documents. For regular documents and folded documents, all of the model's accuracy, precision, recall, and F1-score reached 100%. For damp and creased documents, these metrics ranged from 95.0% to 99.0%. the model's multimodal capability effectively extracts structured information from birth certificates, and its overall performance is notably better than the

traditional OCR baseline models used in previous studies, especially in cases involving damage documents.

This study aims to evaluate the effectivity of GPT-5.5 in official documents degradation handling with various categories. The output of the experiment indicates that, by utilizing the model's strengths in contextual reasoning and multimodal understanding, documents affected by physical damages, shadows, can be processed accurately and consistently extracts information. Though, model's overall consistency decreases due to distortion and heavily crumpled documents. In general, GPT-5.5 proves a robust and adaptable ability to common conditions of the document samples.

Digitizing documents can be done effectively as GPT-5.5 demonstrates a feasible solution for institutional archives. It surpasses the traditional method in providing the accuracy level of information extraction due to its combined capability in context reasoning, semantic understanding, and smart document parsing. It can contribute and improved administrative efficiency and decrease the manual data entry because its context is designed to process local NSO and PSA birth certificates.

Recommendations

The proposed recommendations are as follows based on the findings and conclusions:

1. For future researches, the original dataset can be extended by integrating different types of document conditions including faded and damaged ones, examine the robustness of GPT-5.5.

2. Researchers can integrate geometric correction, perspective transformation, adaptive thresholding, and image restoration algorithms to process heavily creased documents and improve OCR performance.

REFERENCES

1. A. Malladhi, "Transforming information extraction: AI and machine learning in optical character recognition systems and applications across industries," International Journal of Computer Trends and Technology, vol. 71, no. 4, pp. 71–78, 2023.
2. A. Chaudhury et al., "A deep OCR for degraded Bangla documents," ACM Transactions on Asian and Low- Resource Language Information Processing, vol. 21, no. 5, 2022.
3. B. J. Soumya, "Enhancing document image processing: Correcting skew in printed documents using deep learning," Journal of Information Systems Engineering and Management, vol. 10, no. 25s, 2025.
4. W. Liao et al., "DocLayLLM: An efficient multi-modal extension of large language models for text-rich document understanding," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025, pp. 3986–3996
5. Y. K. Chia et al., "M-LongDoc: A benchmark for multimodal super-long document understanding and a retrieval-aware tuning framework," in Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP), 2025, pp. 8341–8363.

6. P. Nitayavardhana et al., “Streamlining data recording through optical character recognition: A prospective multi-center study in intensive care units,” Critical Care, vol. 29, no. 1, 2025.
7. A. Pawar et al., “Automated employee attendance monitoring using liveness face recognition and geofencing in real time,” in Proc. IEEE 5th International Conference on Advances in Electronics, Computers and Communications (ICAECC), 2023, pp. 1–6.
8. C. H. Boe et al., “An automated face detection and recognition for class attendance,” International Journal on Informatics Visualization, vol. 8, no. 3, pp. 1672–1680, 2024.
9. T. Asselborn et al., “Enhancing text recognition of damaged documents through synergistic OCR and large language models,” in Proc. 19th Conference on Computer Science and Intelligence Systems (FedCSIS), 2024, pp. 443–447.
10. M. S. Gambe and F. J. B. Talirongan, “AI-Powered Optical Character Recognition for Automated Timesheet Data Extraction: A Multimodal Approach for Handling Document Degradation,” International Journal of Research and Innovation in Social Science, vol. X, no. III, pp. 2598–2609, 2026.
11. S. A. Singh, “AI-driven document processing: A novel framework for automated invoice data extraction from PDF documents,” International Journal of Multidisciplinary Research, vol. 6, no. 6, pp. 1–8, 2024.
12. G. Abinaya et al., “Automated document processing: Combining OCR and generative AI for efficient text extraction and summarization,” in Proc. International Conference on Smart Electronics and Communication Systems (ICSES), 2024, pp. 1–6.
13. S. K. Abirami et al., “AI-based attendance tracking system using real-time facial recognition,” in Proc. 6th International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2022, pp. 1547–1552.
14. N. Khan et al., “Systematic literature review of machine learning models and applications for text recognition,” IEEE Access, vol. 13, pp. 12838–12860, 2025.
15. Yogish Naik et al., “OCR processing under degraded and distorted document conditions,” 2024.
16. A. M. Tanasa and S. V. Oprea, “Rethinking chart understanding using multimodal large language models,” Computers, Materials & Continua, vol. 84, no. 2, pp. 2475–2492, 2025.
17. A. Nagasubramanian et al., “OCRNet: A robust deep learning framework for alphanumeric character recognition to assist the visually impaired,” Scientific Reports, vol. 15, 2025.
18. W. Wang et al., “Layout-aware multimodal document understanding systems,” 2024.