AI-Powered Optical Character Recognition for Automated Timesheet Data Extraction: A Multimodal Approach for Handling Document Degradation

Mellanie S. Gambe[1], Florence Jean B. Talirongan[2]

[1]Student, Northwestern Mindanao State College of Science and Technology Instructor, St. Peter's College

[2]Professor, Northwestern Mindanao State College of Science and Technology

DOI: https://doi.org/10.47772/IJRISS.2026.100300188

Received: 12 March 2026; Accepted: 17 March 2026; Published: 31 March 2026

ABSTRACT

This study explores the utilization of Optical Character Recognition in extracting employee information from physical datasheets. The development process includes the integration of Google Gemini 2.5 Flash Framework with the implementation of React frontend development. The different states of degradation, namely original, folded, crumpled, and wet, were 20 samples per category for 80 samples. The system achieved high accuracy: 100% accuracy for original documents, 90% for folded documents, 70% for crumpled documents, and 91.66% for wet timesheets, with a final accuracy of 87.92%. This means that context-aware multimodal reasoning is a powerful framework that can substantially reduce the reliance on standard binarization and template-matching in real-life document digitization, achieving 12–47 percentage point higher accuracy than the baseline OCR. This work serves as a baseline in determining document degradation in terms of manual to digital utilization and extraction.

Keywords: multimodal LLM, Gemini 2.5 Flash, optical character recognition, document degradation, structured data extraction, document digitization, prompt engineering, timesheet automation

INTRODUCTION

In today's organizational environment, time management and administrative digitization are crucial to reduce work effort and to decrease administrative burden (Ranjan et al., 2025; Tanasa & Oprea, 2025). Manual processing of timesheets in manufacturing, retail, and service industries in the Philippines is a serious productivity drain on SMEs, with large proportions of them still using paper documentation systems to keep track of operations. Inputting manual data is prone to human error and results in high costs from an administrative standpoint, along with problems of maintaining the integrity of the data (Ranjan et al., 2025). Timesheet records need to be digitized to ensure transparency in organizations, proper payroll documentation for employees, and compliance with labor standards of the Department of Labor and Employment (Malladhi, 2023). Recent AI-based attendance systems using facial recognition and geofencing similarly aim to automate time tracking and reduce manual encoding in organizational settings (Abirami et al., 2022; Pawar et al., 2023; Chauhan et al., 2024; Thohir et al., 2025).

The widely adopted method for converting printed text to machine text is Optical Character Recognition (OCR) (Malladhi, 2023; Singh, 2024). This evolution from simple pattern matching to deep learning and neural networks to interpret context, layout, and meaning (Singh, 2024; Liao et al., 2025) has occurred through the progression of recently evolved technologies in OCR systems. Indeed, these developments revolutionize data capture for domains such as finance and logistics due to the ability to analyze complex documents in real-time (Liao et al., 2025). In contrast, traditional OCR algorithms are characterized by rigid binarization algorithms and template-based extraction methods, which tend to fail on tests of document quality (Malladhi, 2023; Soumya, 2025). Failure to consider any context in depth means they face various issues, such as unbalanced

illumination, variable document construction, or document decay (which frequently occur in practice). Experimental studies have tested the OCR method for document capture, which has achieved great performance in drastically cutting the manual entry and retaining the quality (Nitayavardhana et al., 2025). Here, the possibility of OCR as a real-world alternative to handwritten entry emerges rather than merely a supplement.

From 2025, Multimodal Large Language Models are a new era paradigm in document processing, which aims to perform both visual and linguistic understanding in order to have high accuracy on degraded inputs (Khan et al., 2025; Chia et al., 2025). By directly consuming images and performing context-aware text extraction through prompt engineering, these models circumvent the traditional OCR pipeline completely (Wang et al., 2024). A perennial challenge of research is that traditional OCR systems often break down with the presence of physical degradation, such as creases, water stains, or uneven illumination, affecting documents (Asselborn et al., 2024; Nagasubramanian et al., 2025). Although these engines have become popular with the advent of clean scans, they perform poorly if the original scans are compromised, and human intervention is costly to correct (Chaudhury et al., 2022; Yogish Naik et al., 2024).

This study proposed and compared a computer application using the Google Gemini 2.5 Flash multimodal model to automate the extraction of three vital fields (Employee Number, Name, Department) from degraded timesheets. The novelty comes from switching from rule-based character recognition to the context-aware multimodal reasoning approach, which forms a robust framework that avoids the disadvantages of conventional binarization and is compatible with enterprise web environments (Tanasa & Oprea, 2025; Liao et al., 2025; Wang et al., 2024). In parallel, educational and enterprise systems have begun adopting automated attendance tracking through face detection and recognition to minimize supervisor burden and improve record accuracy (Chaudhury et al., 2022; Boe et al., 2024).

LITERATURE REVIEW

Traditional OCR Limitations in Complex Environments

Traditional OCR engines, being computationally straightforward, do not inherently understand context or semantic meaning (Singh, 2024). Based on previous work, traditional OCR performance has shown huge importance on the sharp character borders and high-quality scans, with high accuracy falling off sharply when the input images contain noise, blur, or geometric distortion (Chaudhury et al., 2022; Yogish Naik et al., 2024; Khan et al., 2025). Systems such as Tesseract tend to mess up the reading order of intricate layouts, and the resulting output is typically subject to an extensive post-processing process before it can be used for structured data extraction (Malladhi, 2023; Singh, 2024). Template-specific extraction techniques are intrinsically brittle since it is still necessary to manually tweak the format of their code for even a slight change in the structure or font characteristics of a document (Singh, 2024). Conventional software engines do not 'learn' according to the document context, thus heavy pre-processing is required to prepare images, as well as skew correction and noise reduction (Yogish Naik et al., 2024; Soumya, 2025; Asselborn et al., 2024). The computation requirements involved in preprocessing may surpass the cost of direct multimodal inference, which is the problem behind traditional OCR in inefficient cost accounting in degraded document batches (Tanasa & Oprea, 2025; Liao et al., 2025).

Impact of Physical Degradation on Data Integrity

Physical damage also generates visual "noise" which hampers automated extraction pipelines and propagation further down the hierarchy (Asselborn et al., 2024). Moisture damage and smudging generate inconsistent lighting and ink bleeding that cannot be mitigated by conventional binarization algorithms, which assume a clean binary image (black-white) character separation (Chaudhury et al., 2022). Moreover, degraded documents tend to involve resource-intensive restoring tasks, such as image reconstruction algorithms, dilation/erosion filters, and multi-scale processing (Soumya, 2025; Asselborn et al., 2024). The net consequence is that batch processing of degraded documents becomes technically challenging for organizations, especially in resource-constrained settings like the Philippines (Tsai & Li, 2023).

Layout-Aware Multimodal Document Understanding as Advanced OCR Paradigm

Due to contemporary document processing moving beyond pixel-based OCR to layout-aware multimodal methods with a multimodal approach in mind, the researchers incorporated visual, textual, and spatial information simultaneously (Liao et al., 2025; Wang et al., 2024). According to Wang et al. (2024), layout-aware models combine bounding box information, visual tokens, and spatial-aware attention to identify and manipulate the structure of a document (including the relationships of regions, the order of reading, and the disposition of content) for higher performance in document intelligence tasks (e.g., in form extraction, table comprehension, and structured data recovery) (Liao et al., 2025; Wang et al., 2024). It is especially the case of varying layouts that are often seen in real documents: forms with varied positions in the field, invoices nested in tables, and timesheets formatted for non-standard formats (Liao et al., 2025). In contrast to traditional binarization-based algorithms, layout-aware multimodal models efficiently manage mixed-media documents while attaining leading performance across multi-content types and avoiding individual pipeline work (Tanasa & Oprea, 2025; Singh, 2024; Liao et al., 2025; Wang et al., 2024).

Nevertheless, the problem of extreme physical degradation remains: layout-aware multimodal models can hardly extract this useful information in the face of extreme physical degradation wherein the visual cues are badly corrupted due to document creases, aggressive crumpling, or in the case of severe moisture damage since spatial information alone cannot extract the model (Yogish Naik et al., 2024; Soumya, 2025; Asselborn et al., 2024). In light of this, this paper takes advantage of the extra aspect of linguistic reasoning, in that multimodal Large Language Models using visual perception and semantic understanding are being employed that can infer missing or distorted information from the surrounding context (Tanasa & Oprea, 2025; Liao et al., 2025; Wang et al., 2024). This way of thinking aims to move beyond purely vision-based layout understandings to hybrid vision-language reasoning, which can recover incoming information even where visual cues are not sufficient alone, which is a key step in more resilient document processing for real-world degraded inputs (Liao et al., 2025; Wang et al., 2024; Asselborn et al., 2024).

AI-Based Attendance Automation Systems and Multimodal Recognition

Apart from document-oriented OCR applications, a lot of work was done to automate attendance and employee time tracking using AI-powered recognition systems. Previous work has shown real-time tracking of attendance via deep-learning-based facial recognition (Abirami et al., 2022; Boe et al., 2024), a system that combines liveness detection and geofencing to mitigate proxy attendance (Pawar et al., 2023; Chauhan et al., 2024), and web-based employee attendance portals that implement face recognition in browser environments (Thohir et al., 2025). These systems are mainly adapted to clean visual inputs that are recorded directly through a camera or mobile device. The present work deals with a more challenging problem: to recover structured employee timesheet information from physically degraded paper records through multimodal OCR.

METHODOLOGY

This study adopts the DTM-OCR (Degraded Timesheet Multimodal OCR) model architecture shown in Figure 1, which integrates a React-based frontend, a dual-prompt Gemini 2.5 Flash extraction engine, and a validation-verification layer.

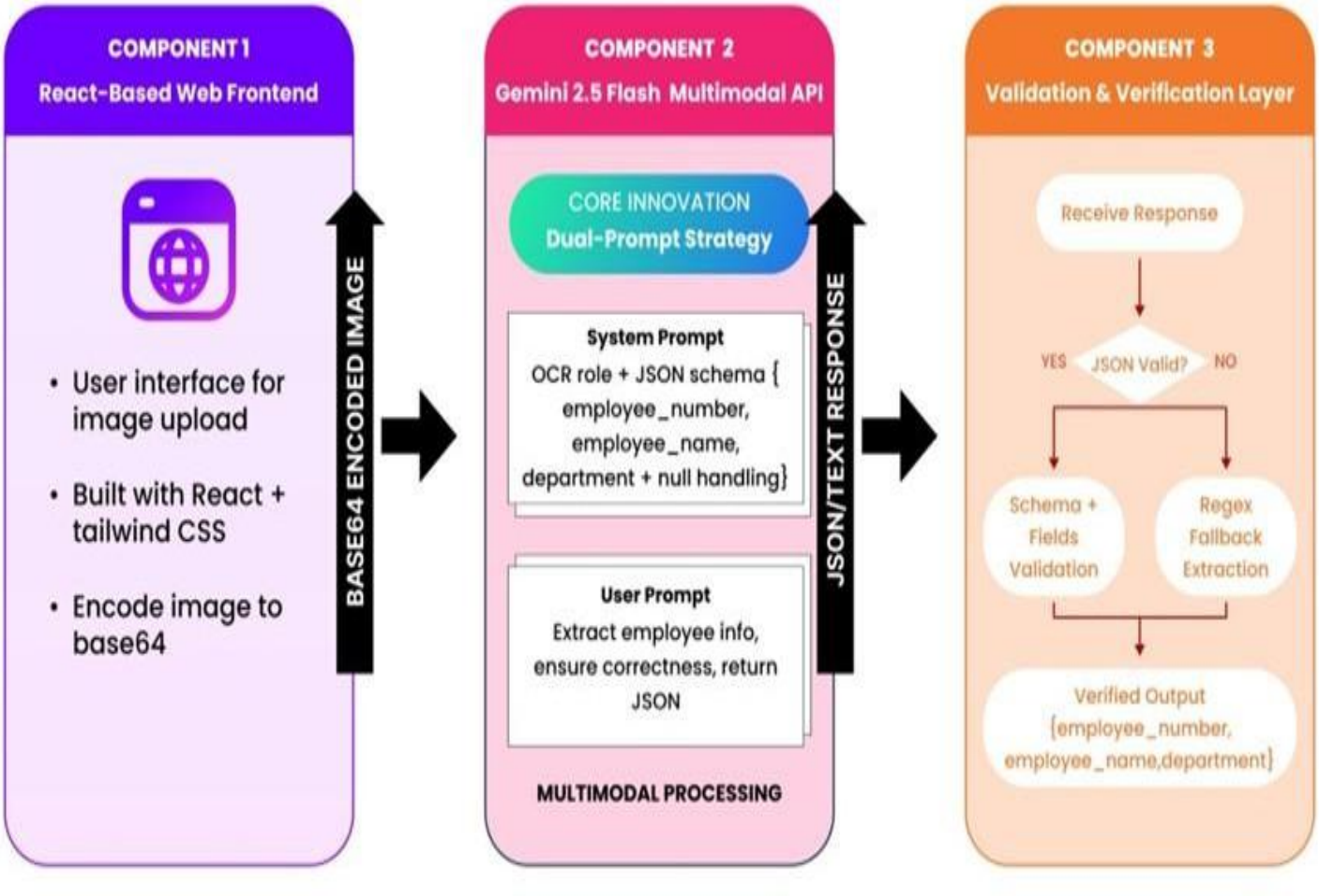

Figure 1. Proposed DTM-OCR model architecture

The introduction of this model yields the top-level framework which frames the complete system and the subsequent subsections provide detailed evidence of how the implemented components are handled, how the dataset is defined, and how the extraction performance has been examined, based on the degradation conditions.

Materials and Study Design

The study was conducted at St. Peter's College, Iligan City, Philippines. Testing: There was actual testing of 80 physical timesheets sourced from the institutional administrative archive (fiscal year 2024--2025) as real life records of the employees. Samples were systematically classified into four environmental degradation conditions: (1) Original (pristine intact timesheets stored in controlled climate conditions) (n = 20); (2) Folded (timesheets exposed to single or multiple folds resulting in characteristic crease shadow formations) (n = 20); (3) Crumpled (aggressively crumpled sheets that have recognizable three-dimensional surface deformation) (n = 20); (4) Wet (timesheets left in water to mimic workplace spills with natural drying) (n = 20). This experimental nature is consistent with previous research on intelligent attendance monitoring programs that fuse automated identification with real-world deployment constraints (Tsai & Li, 2023).

System Architecture and Data Flow

The system architecture consists of 3 primary components, namely, (1) React-based web frontend for user interaction and image submission, (2) Google Gemini 2.5 Flash multimodal API for image analysis and structured data extraction, and (3) the JSON validation and output verification layer (Tsai & Li, 2023; Liao et al., 2025; Thohir et al., 2025). The researchers designed an appealing React frontend that integrated timesheet images with custom Tailwind CSS. Validated images were encoded into Base64 and forwarded to the Gemini 2.5 Flash API, where structured system and user prompts were used for multimodal processing.

In Figure 2, the web-based Time Card Extractor homepage presents a simple upload interface where users submit timesheet images for processing. The drag-and-drop area standardizes image intake, ensuring that all samples follow the same preprocessing and extraction pipeline.

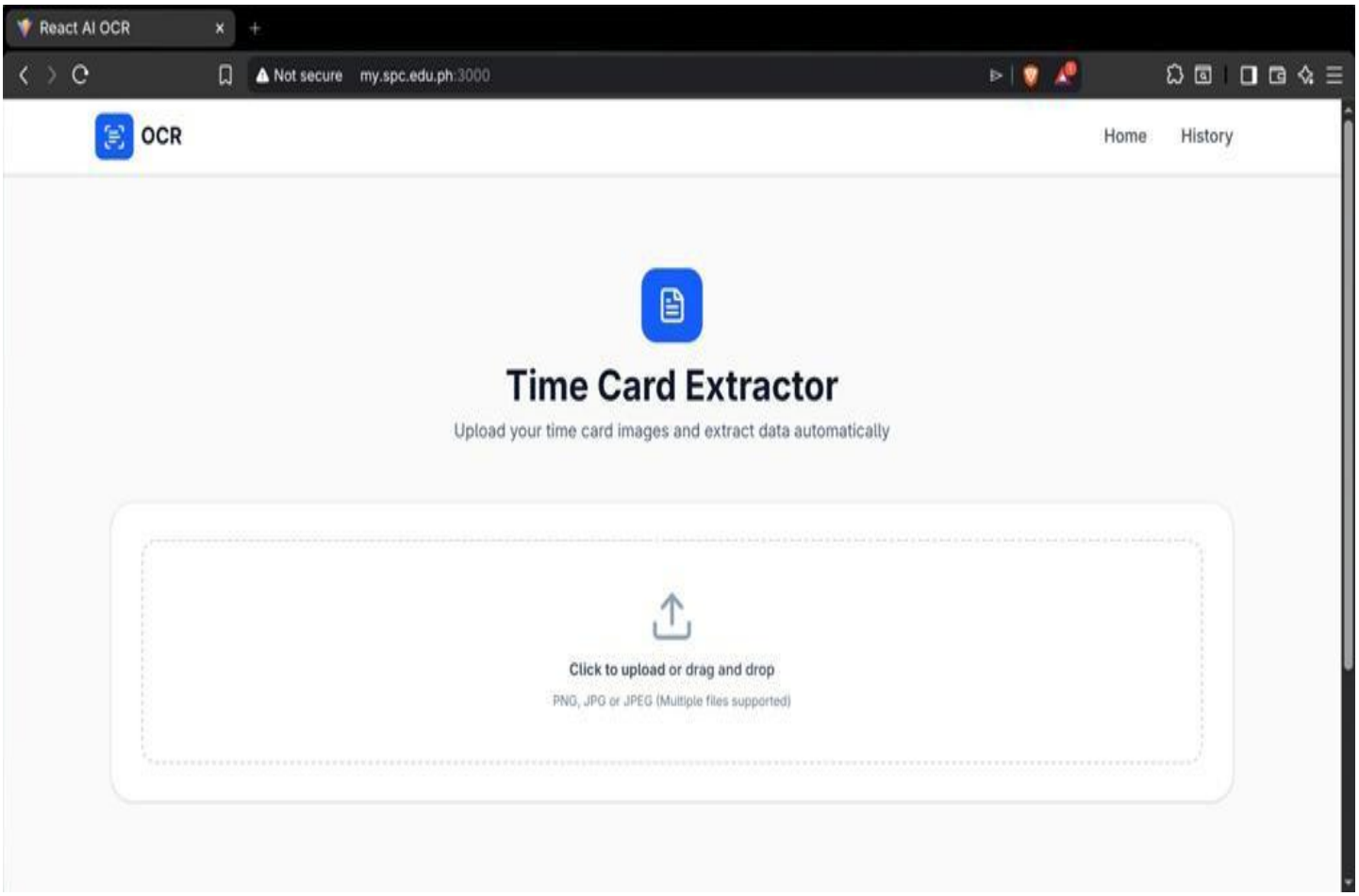

Figure 2. Homepage upload interface

Structured Prompt Engineering

The system prompt for Gemini 2.5 Flash stated: "You are a precision OCR engine to pull out employee timesheet data. Gather and retrieve merely employee data for the image provided: Employee Number, Employee Name, and Department. The output MUST contain valid JSON, with keys employee_number, employee_name, and department. Use null if extraction cannot be confident." Extracting from the user prompt: "You get info about the employee from the timesheet image and return it as a JSON object. Be rigorous about correctness. If not sure you should return null in that field." This double-prompt method mitigated model error (with respect to hallucination), and led to a deterministic, user-friendly output easily injected into payroll systems in the future (Liao et al., 2025; Wang et al., 2024).

Extraction Algorithm (Pseudocode)

Algorithm: MultimodalOCRExtraction

Input: timesheet_image (binary image data)

Output: extracted_fields (JSON object with employee_number, employee_name, department)

Begin

1. Encode image to Base64 format

2. Prepare API request with system_instruction and user_prompt

3. Call Gemini 2.5 Flash API with {image, system_instruction, user_prompt}

4. Receive response from API (text or JSON)

5. Attempt JSON parsing

IF valid JSON THEN

5a. Validate field types (string or null)

5b. Return extracted_fields = {employee_number, employee_name, department} ELSE

5c. Apply regex extraction on response text

5d. Build extracted_fields = {employee_number, employee_name, department} from regex matches

6. Return extracted_fields

End

Testing Protocol and Data Collection

The researchers conducted a standard seven-step manual evaluation with our 80 timesheet samples. First, each timesheet was photographed on a smartphone camera in the office with controlled lighting and exported as a JPEG image at 72 DPI for uploading to the Gemini 2.5 Flash API. The picture was then uploaded to the API along with the structured system and user prompts, and the JSON response that has the three extracted data fields - employee_number, employee_name, department with the timestamp of each sample was recorded and logged. Figure 3 shows the main extraction view that shows the raw time card alongside parsed employee data. It provides immediate visual representation for the parsed JSON fields with the source document. As illustrated in Figure 4, time card images and resultant outputs aggregate into a history screen that provides historical extractions. This method assists auditing and error analysis as it allows a reviewer to come back to see any extraction instance and both raw image and extracted files.

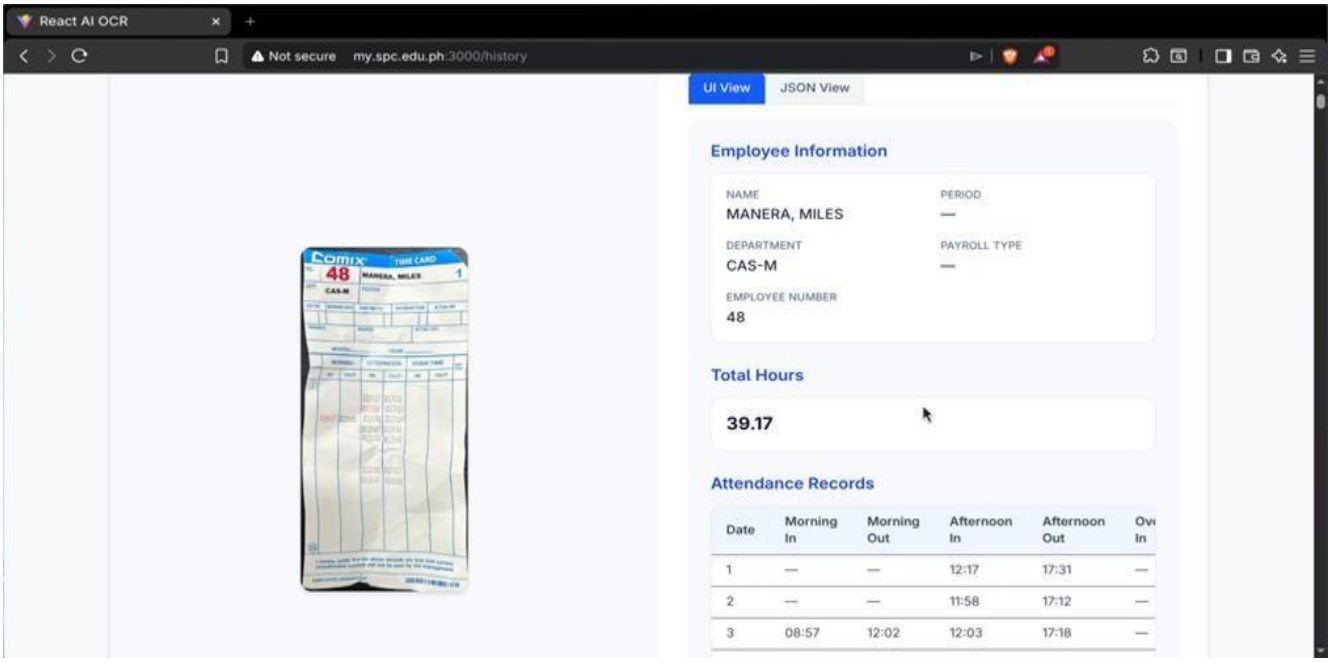

Figure 3. Single extraction UI (one card + extracted data)

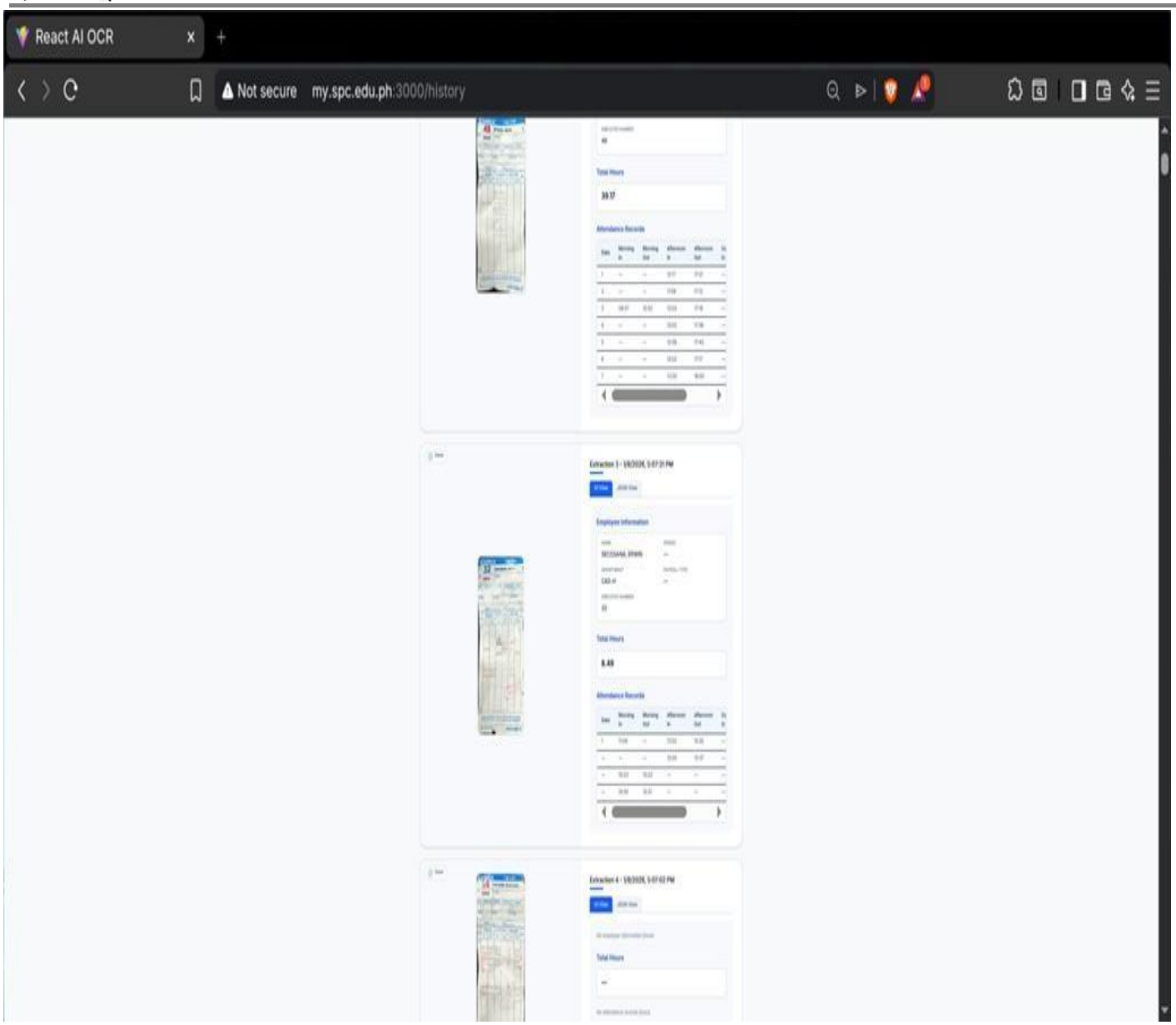

Figure 4. Extraction history list

Third, the extracted values for the three target fields were manually verified by the researchers through character-by-character comparison against the ground truth values printed on the original timesheet. For every sample and for each field, the result was manually marked as correct (checkmark) if it matched exactly, or incorrect (x) if any discrepancy was observed, using case-sensitive comparison for names and digit-perfect comparison for employee numbers. Finally, all accuracy metrics were computed manually using a spreadsheet. For each document condition (original, folded, crumpled, wet), the number of correctly extracted fields was divided by the total number of fields and expressed as a percentage. These percentages were then averaged across the three fields to obtain the overall accuracy for that condition. The reported overall system accuracy of 87.92% corresponds to a manually calculated Exact Match Rate (EMR), defined as the proportion of fields whose extracted values exactly match their ground truth counterparts (Liao et al., 2025; Soumya, 2025; Khan et al., 2025).

RESULTS AND DISCUSSION

Overall System Performance by Document Condition

The table in this paper shows timesheet extraction accuracy results depending on the document degradation condition for 80 timesheets.

Table 1. Timesheet Extraction Accuracy Results by Document Degradation Condition

Condition	Employee Number	Employee Name	Department	Overall Accuracy
Original (n=20)	100% (20/20)	100% (20/20)	100% (20/20)	100%
Folded (n=20)	90% (18/20)	95% (19/20)	85% (17/20)	90%
Crumpled (n=20)	70% (14/20)	70% (14/20)	70% (14/20)	70%
Wet (n=20)	85% (17/20)	100% (20/20)	90% (18/20)	91.66%
Overall (n=80)	**86.25%**	**91.25%**	**86.25%**	**87.92%**

Table 1 presents core features, which show the multimodal OCR performance under four conditions of documents. Finally, in the final analysis, all three fields of the original timesheets have all accuracy up to 100%, meaning the prompt engineering along with connecting the API would be correct if everything was just perfect (and that is the baseline on which to judge our accuracy). The folded documents, having some crease-induced shadows, show mostly 90% accuracy overall, but the missing characters were mainly inferred from a context-sensitive approach in which the model deduced most accurately. Most importantly, crumpled documents give you the lowest accuracy of 70% of the models since the deep three-dimensional surface deformations and complex patterns overlap so extensively that they can fragment character boundaries into an overlap. The result for wet documents showed overall performance of 91.66% (second to pristine originals), 100% Employee Name extraction quality (which means the model is able to handle the impact of ink diffusion and color changes on longer text sequences with semantic redundancy). In this way, an excellent realistic degree of system accuracy (Tanasa & Oprea, 2025; Liao et al., 2025; Asselborn et al., 2024); 87.92%, indicates considerable potential for deployment in the field, regardless as to the existence of degraded documents.

Field-Level Extraction Performance Analysis

Field-level performance data was analyzed utilizing Exact Match Rate (EMR) across all 240 field extractions (80 samples × 3 fields). Table 2 gives detailed sample group and error type performance metrics. Detailed performance (Table 2) for each extraction is reported in 240 records for fields from 80 samples × 3 fields. This very extreme condition-specific bias has the effect of explaining an EMR of 87.92% (211/240 succeeds). The distribution of failure is very skewed such that crumpled files contribute almost 62% (18 of 29 failed fields), which is that geometric distortion is an important technical challenge. Wet documents meanwhile only suffer a failure rate of 8.33% even when they are visually deteriorated. The occurrence of 10% failure when the documents are folded lies between the two extremes, reflecting moderate difficulty by itself or somewhat influenced by situation-dependent inferences. That distribution shows that preprocessing is a focus point: crumpled documents could be preprocessed for the maximum amount of work (e.g., geometric correction) and provide the best ROI (Soumya, 2025; Asselborn et al., 2024).

Table 2. Field-Level Exact Match Rate (EMR) Analysis across Document Conditions

Sample Group	Total Samples	Total Fields	Successful	Failed	Primary Error Cause	Failure Rate
Original	20	60	60	0	N/A	0%

Folded	20	60	54	6	Crease shadows (50%), text occlusion (50%)	10%
Crumpled	20	60	42	18	Severe geometric distortion (100%)	30%
Wet	20	60	55	5	Ink diffusion on numerics (60%), ink dissolution (40%)	8.33%
Total	**80**	**240**	**211**	**29**	**Overall EMR: 87.92%**	**12.08%**

Critical field-level vulnerability analysis reveals:

Employee Name (91.25% accuracy): Overall best performance when putting context behind long character-level sequences. Stably reached 100% accuracy on wet papers despite the presence of ink diffusion (Tanasa & Oprea, 2025; Asselborn et al., 2024).

Employee Number (86.25% accuracy): Intermediate susceptibility as 4 to 6 digit sequences isolation lacks contextual integrity, rendering pixel-level degradation. Water damage resulted in confusion of digits (8 to 3, 5 to 6) (Chaudhury et al., 2022; Asselborn et al., 2024).

Department (86.25% accuracy): Intermediate; small strings of text in text form mean less context for error recovery. A breakdown was reported in folded and crumpled documents (Yogish Naik et al., 2024; Asselborn et al., 2024).

Performance Comparison of Multimodal LLM with Traditional OCR

Previous experimental studies with degraded documents have repeatedly found that conventional OCR achieves 40--75% accuracy on degraded documents (Chaudhury et al., 2022; Yogish Naik et al., 2024; Khan et al., 2025; Asselborn et al., 2024; Nagasubramanian et al., 2025). Indeed, the accuracy score of Tesseract OCR is only 63.33% for 60 Philippine official documents in nature, as noted in a recent study (Abinaya et al., 2024). But these figures differ widely for the true text of an application: the study reports a 24.6 and 12--47 percentage point improvement over literature baseline measurements. Related to the successful deployment of OCR systems in intensive care units to alleviate the burden of manual data recording (Nitayavardhana et al., 2025), the results suggest that even less than ideal AI-based extraction can considerably reduce the administrative burden in timesheet management.

Table 3. Performance Comparison: Multimodal LLM (Gemini 2.5 Flash) versus Traditional OCR Baselines.

Degradation Scenario	**Traditional OCR**	**Gemini 2.5 Flash**	**Improvement (pp)**	**Sources**
Overall Real-world degraded	40--75%	87.92%	+12 to +47	(Chaudhury et al., 2022; Yogish Naik et al., 2024; Khan et al., 2025; Asselborn et al., 2024; Nagasubramanian et al., 2025)

Original/High-quality	95%+	100%	Comparable/Superior	(Chaudhury et al., 2022; Yogish Naik et al., 2024; Khan et al., 2025)
Folded/Moderate degradation	60--75%	90%	+15--30	(Chaudhury et al., 2022; Yogish Naik et al., 2024; Asselborn et al., 2024)
Crumpled/Severe degradation	40--55%	70%	+15--30	(Chaudhury et al., 2022; Yogish Naik et al., 2024; Asselborn et al., 2024)
Wet/Environmental damage	50--65%	91.66%	+26--41	(Chaudhury et al., 2022; Yogish Naik et al., 2024; Asselborn et al., 2024)
Philippine Real-world (Tesseract)	63.33%	87.92%	+24.59	(Abinaya et al., 2024)

As Table 3 illustrates, the results of this study are placed in the context of the larger ecosystem of OCR research, using Gemini 2.5 Flash performance, comparing it with the known baseline performance of typical and earlier deep learning OCR systems. On original/high-quality documents, both approaches show comparable performance (95%+ versus 100%), supporting the argument that multimodal LLMs do not sacrifice performance on clean inputs. The significant discrepancy, however, can be seen on degraded documents: folded scenarios show 15--30 percentage-point gains (+15--30pp), crumpled scenarios show comparable advantages (+15--30pp), and wet scenarios show the greatest change (+26--41pp). This 24.59 percentage-point gain over the recent Tesseract benchmark on real-world Philippine documents (Abinaya et al., 2024) provides straightforward, country-targeted evidence that multimodal techniques are indeed a significant improvement. The relative advantage of such a paradigm to digitize documents from pixel-level character recognition to context-aware linguistic reasoning has thus been confirmed (Tanasa & Oprea, 2025; Chaudhury et al., 2022; Yogish Naik et al., 2024; Liao et al., 2025; Asselborn et al., 2024; Nagasubramanian et al., 2025) in this comparative study.

These results establish a new empirical baseline for multimodal LLM-based OCR on degraded Philippine timesheet records, demonstrating substantial and consistent gains over traditional OCR methods reported in the literature.

CONCLUSIONS AND RECOMMENDATIONS

Key Findings

The authors have shown that AI-assisted OCR using multimodal LLMs should be practical and inexpensive to facilitate the digitization of damaged employee records relevant to the Philippine workplace. In this project, the React frontend, structured prompt engineering and Google Gemini 2.5 Flash achieve 87.92% accuracy across 80 degraded real-world timesheets.

Key findings include:

1. Multimodal reasoning eliminates the limitations of OCR in traditional systems: the system had 70% accuracy for crumpled timesheets, which is better than the traditional models, given the situational sensitivity.

2. Wet documents achieved 91.66% overall accuracy, with 100% accuracy for Employee Name fields, indicating that the multimodal model can exploit contextual redundancy in longer text sequences.
3. Crumpling represents the key failure mode, as a 30% field failure rate (field failure rate of crumpled records) demonstrates that the high geometric distortion in papers due to crumpling surpasses the preprocessing capabilities of the process.
4. Structured prompt engineering ensures deterministic responses: limiting responses to JSON format minimizes the risks of hallucination, and integration into downstream systems is straightforward.
5. Transition from rule-based to context-aware reasoning: This study demonstrates that applying linguistic reasoning to document images is superior to algorithmic character recognition.

The research fills the gap observed in the resilient OCR frameworks related to the actual conditions in Philippine workplaces.

RECOMMENDATION

The study needs to be carried out to optimize image preprocessing for crumpled documents using perspective correction algorithms and histogram equalization for future work. Moreover, multi-page document batching achieves batch processing capability for 50--100 timesheets per API call, which allows saving per-document costs and increases throughput.

REFERENCES

1. Abinaya, G., Aparna, K. H., Keerthika, M., Harshini, S. R., & Jothi, K. R. (2024). Automated document processing: Combining OCR and generative AI for efficient text extraction and summarization. In Proceedings of the 2024 International Conference on Smart Electronics and Communication Systems (ICSES) (pp. 1--6). IEEE. https://doi.org/10.1109/ICSES63760.2024.10910510
2. Abirami, S. K., Jyothikamalesh, S., Sowmiya, M., Abirami, S., Mary, S. A. L., & Jayasudha, C. (2022). AI-based attendance tracking system using real-time facial recognition. In 6th International Conference on Electronics, Communication and Aerospace Technology (ICECA 2022) - Proceedings (pp. 1547--1552). IEEE. https://doi.org/10.1109/ICECA55336.2022.10009331
3. Asselborn, T., Dorokhova, M., Šafránek, D., Šimečková, P., & Vepřek, V. (2024). Enhancing text recognition of damaged documents through synergistic OCR and large language models. In M. Ganzha, L. Maciaszek, M. Paprzycki, & D. Ślęzak (Eds.), Proceedings of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS) (pp. 443--447). IEEE. https://doi.org/10.15439/2024f7400
4. Boe, C. H., Ng, K. W., Haw, S. C., Naveen, P., & Anaam, E. A. (2024). An automated face detection and recognition for class attendance. International Journal on Informatics Visualization, 8(3), 1672--1680. https://doi.org/10.62527/joiv.8.3.2967
5. Chaudhury, A., Mukherjee, P. S., Das, S., Biswas, C., & Bhattacharya, U. (2022). A deep OCR for degraded Bangla documents. ACM Transactions on Asian and Low-Resource Language Information Processing, 21(5), Article 91. https://doi.org/10.1145/3511807
6. Chauhan, V., Singh, H., Dewari, K., & Kumar, I. (2024). Efficient employee tracking with smart attendance system using advanced face recognition and geofencing. In 2nd International Conference on Sustainable Computing and Smart Systems (ICSCSS 2024) - Proceedings (pp. 1214--1219). IEEE. https://doi.org/10.1109/ICSCSS60660.2024.10625038
7. Chia, Y. K., Xu, P., Xie, S. M., Rudzicz, F., & Kawaguchi, K. (2025). M-LongDoc: A benchmark for multimodal super-long document understanding and a retrieval-aware tuning framework. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 8341--8363). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.emnlp-main.469
8. Khan, N., Din, S., Rehman, S. U., Hameed, A. A., Qureshi, S. S., Algarni, A. D., Shah, H., & Elmannai, H. (2025). Systematic literature review of machine learning models and applications for text recognition. IEEE Access, 13, 12838--12860. https://doi.org/10.1109/ACCESS.2025.3618109
9. Liao, W., Xu, Y., Li, H., Wang, L., Zeng, J., Zhong, W., Chen, G., Guo, S., Zhang, S., Zhang, K., Liu, L., Liu, Z., & Sun, M. (2025). DocLayLLM: An efficient multi-modal extension of large language models for text-rich document understanding. In Proceedings of the IEEE/CVF Conference on Computer Vision

and Pattern Recognition (CVPR) (pp. 3986--3996). IEEE. https://doi.org/10.1109/cvpr52734.2025.00382

10. Malladhi, A. (2023). Transforming information extraction: AI and machine learning in optical character recognition systems and applications across industries. International Journal of Computer Trends and Technology, 71(4), 71--78. https://doi.org/10.14445/22312803/ijctt-v71i4p110
11. Nagasubramanian, A., Srinivasan, G. K., Sharma, A., Krishnamurthy, B., & Madhan Kumar, S. (2025). OCRNet: A robust deep learning framework for alphanumeric character recognition to assist the visually impaired. Scientific Reports, 15, Article 1847. https://doi.org/10.1038/s41598-025-25278-9
12. Nitayavardhana, P., Rattanathananon, P., Chinswang, S., Lekhakul, A., Charoenphon, C., Tulyathan, T., Saksirinukul, A., & Kiatsopit, K. (2025). Streamlining data recording through optical character recognition: A prospective multi-center study in intensive care units. Critical Care, 29(1), Article 88. https://doi.org/10.1186/s13054-025-05347-1
13. Pawar, A., Hiwanj, R., Koparde, P., Chikmurge, D., & Barve, S. (2023). Automated employee attendance monitoring using liveness face recognition and geofencing in real time. In Proceedings of the IEEE 2023 5th International Conference on Advances in Electronics, Computers and Communications (ICAECC 2023) (pp. 1--6). IEEE. https://doi.org/10.1109/ICAECC59324.2023.10560301
14. Ranjan, R., Singh, R., Kumar, J., & Tripathi, S. (2025). Time management for leaders and impact on productivity: A review study. International Journal of Innovative Research in Science & Studies, 8(2), 180--185. https://doi.org/10.53894/ijirss.v8i2.5286
15. Singh, S. A. (2024). AI-driven document processing: A novel framework for automated invoice data extraction from PDF documents. International Journal of Multidisciplinary Research, 6(6), 1--8. https://doi.org/10.36948/ijfmr.2024.v06i06.32247
16. Soumya, B. J. (2025). Enhancing document image processing: Correcting skew in printed documents using deep learning. Journal of Information Systems Engineering and Management, 10(25s), Article s4090. https://doi.org/10.52783/jisem.v10i25s.4090
17. Tanasa, A. M., & Oprea, S. V. (2025). Rethinking chart understanding using multimodal large language models. Computers, Materials & Continua, 84(2), 2475--2492. https://doi.org/10.32604/cmc.2025.065421
18. Thohir, M. I., Kharisma, I. L., & Ika. (2025). Web-based employee attendance system utilizing face recognition and CNN via Face-API.js. Bit-Tech, 8(2), 390--400. https://doi.org/10.32877/bt.v8i2.2828
19. Tsai, M. F., & Li, M. H. (2023). Intelligent attendance monitoring system with spatio-temporal human action recognition. Soft Computing, 27(8), 4517--4531. https://doi.org/10.1007/s00500-022-07582-y
20. Wang, D., Bai, J., Zheng, B., Lin, J., Zhan, J., Fei, L., Li, L., Li, Y., Zhang, X., Li, Y., Zhou, J., Babaev, A., Yu, Y., Tiwari, A., & Kar, A. (2024). DocLLM: A layout-aware generative language model for multimodal document understanding. In L. Ku, A. Martins, & V. Srikumar (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL) (Volume 1: Long Papers) (pp. 8541--8567). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.463
21. Yogish Naik, G. R., Shashidhara, B., & Amith, G. K. (2024). A review on text extraction techniques for degraded historical document images. In 2nd IEEE International Conference on Advances in Information Technology (ICAIT) (pp. 1--6). IEEE. https://doi.org/10.1109/ICAIT61638.2024.10690761