Automated Data Extraction from Aircraft Fuel Invoices Using PaddleOCR

Reinaldy Hutapea[1], Vanessa Harwanto[2], and Samuel Situmeang[1]
[1]Faculty of Informatics and Electrical Engineering, Institut Teknologi Del, North Sumatra, Indonesia
[2]Faculty of Master of Digital Economy, Binus University, Jakarta, Indonesia

Corresponding author: Samuel Situmeang (email: samuel.situmeang@del.ac.id).

ABSTRACT This study presents an automated data extraction system for aircraft fuel invoice documents using PaddleOCR, a deep learning-based optical character recognition (OCR) technology. The system is designed to address the challenges of extracting information from complex and unstructured document formats, which traditionally require extensive manual processing. To enhance performance, the system incorporates image pre-processing techniques and artificial intelligence-based validation methods, ensuring higher accuracy in recognizing aviation-specific details such as flight identifiers and fuel data. Evaluation of the system demonstrates notable improvements in both time efficiency and accuracy. On average, documents can be processed in under 60 seconds with high recognition rates for clean, standard-quality inputs. While performance decreases with noisy or small-text documents, results indicate that accuracy can be further improved through deep learning-based denoising and training with aviation-specific datasets. The system also proves scalable, successfully handling up to 640 documents without compromising performance, suggesting its feasibility for large-scale industrial deployment. Beyond technical efficiency, the system delivers tangible economic benefits by reducing operational costs, minimizing transaction discrepancies, and enabling staff to focus on higher-value strategic tasks. Furthermore, it establishes a foundation for future enhancements, including integration with ERP systems, multilingual OCR support, and handwriting recognition. Overall, this research highlights the potential of PaddleOCR-based automation to significantly transform document management in the aviation industry and offers promising opportunities for adoption across other data-intensive sectors.

KEYWORDS Automated Data Extraction, Aviation Industry, Invoice, PaddleOCR

I. INTRODUCTION

In the rapidly developing digital era, operational efficiency is key to maintaining competitiveness and increasing productivity in various industries [1], [2], [3]. Document processing automation is critical to operational efficiency, especially in invoice management and financial transactions [4], [5]. The aviation industry, which relies on processing large amounts of invoices, still faces challenges in data processing speed, accuracy, and scalability [6]. In 2023, the global market size of the airline industry was valued at 762.8 billion U.S. dollars, highlighting the massive volume of transactions that need to be efficiently processed. Despite the growth, manual processes remain widespread, putting the industry at risk of errors and delays in decision-making. According to recent statistics, 66% of businesses still use Excel spreadsheets for invoice processing, and 38% rely on even less efficient methods, such as emails and whiteboards [7]. These methods are prone to human error, which significantly hampers operational efficiency.

Despite the growing trend in adoption of automation, the transition to fully automated document processing is still ongoing. The global market penetration of document processing software has been steadily increasing, with 58.4% of businesses utilizing such software in 2021. This percentage grew to 64.46% in 2023 and is projected to reach 62.05% by 2024 [8]. This rise in adoption reflects the growing recognition of the need for automation across industries [9], [10] . However, the adoption rate still shows significant room for improvement, particularly in sectors like aviation, where the volume of financial transactions and invoices is enormous. The continued reliance on manual methods underscores the need for more advanced solutions to manage the increasing transaction load.

A 2021 survey highlighted the top benefits of using automation, with 66% of respondents citing the reduction of risks related to performance issues, equipment failure, data breaches, and compliance violations as the primary advantage of automation. Additionally, 50% of respondents noted that automation allows IT staff to focus on strategic initiatives, driving innovation and business growth [11]. These benefits underscore the broader impact of automation,

particularly in industries like aviation, where operational efficiency and data accuracy are paramount.

Researchers and companies have widely used Optical Character Recognition (OCR) to automate data extraction from physical and digital documents [12], [13], [14], [15], [16], [17]. However, conventional OCR methods such as Tesseract often have limitations in handling small text, low image quality, and complex invoice layouts [18], [19], [20]. These limitations lead to information extraction errors especially in documents with various non-standard formats. PaddleOCR[1], a deep learning-based OCR system, is proposed to improve the extraction accuracy in aviation invoices to solve these challenges. PaddleOCR has the advantage of detecting text with high precision and handle more complex document layout variations compared to traditional OCR methods [21].

This research contributes to the field by developing an automation system based on PaddleOCR to enhance efficiency, accuracy, and processing speed of invoice handling in the aviation industry. By focusing on PaddleOCR, the system addresses key challenges in invoice processing, reducing reliance on manual methods, improving data accuracy, and accelerating operational decision-making. We evaluate the system through key metrics, including text extraction accuracy, processing time, and robustness to varying document conditions such as low lighting and high noise. The findings of this study not only contribute to improving automation in the aviation industry but offer insights into applying similar systems in other industrial sectors that require large-scale document automation.

II. METHODS

A. STREAMLINING TEXT EXTRACTION FROM AIRCRAFT FUEL INVOICES

OpenCV (Open Source Computer Vision Library) is a widely used open-source software library for image processing and computer vision [22]. OpenCV provides various functions for image analysis, such as color conversion, noise removal, and morphology management. The library supports several programming languages, including C++, Python, and Java, making it a flexible tool for a variety of computer vision applications.

In this study, OpenCV is used in the pre-processing Stage of invoice document images to prepare the image to be more easily recognized before the text is extracted with PaddleOCR [23].

The pre-processing flow, as illustrated in Figure 1, begins with converting the image to grayscale (Grayscale) to simplify color information and reduce computational complexity. Next, a denoising step is applied to eliminate noise and artifacts that could interfere with text recognition. To enhance contrast and improve text visibility, Contrast-Limited Adaptive Histogram Equalization (CLAHE) is implemented, allowing better differentiation between text and background. The image is then processed using thresholding, which converts it into a binary format, making text boundaries clearer. Subsequently, morphology operations refine character structures and remove unwanted elements. These pre-processing steps ensure that the document images are optimized for accurate text extraction using PaddleOCR.

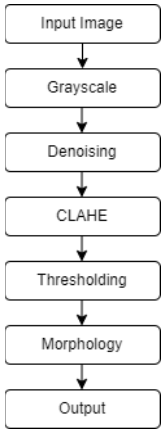


Figure 1. Document Pre-processing Flow for Aircraft Fuel Invoices

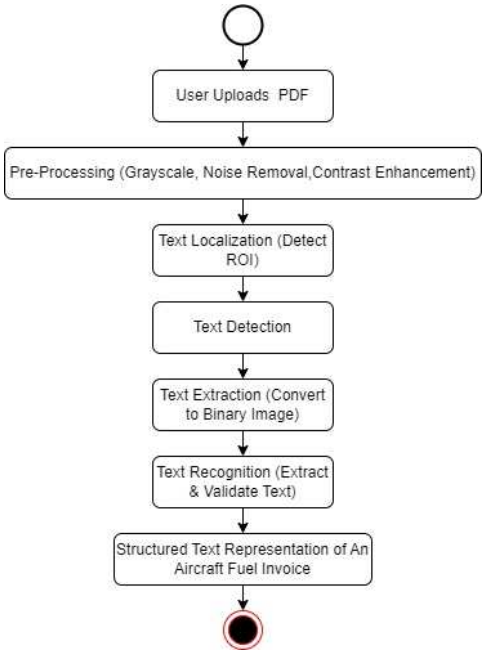


Figure 2. Methodology Flow for Automated Invoice Data Extraction

The research methodology process consists of a series of steps aimed at automating the extraction of important information from uploaded invoice documents. Figure 2 illustrates the step-by-step methodology applied in this process. The following is the workflow methodology applied in this study:

[1] https://github.com/PaddlePaddle/PaddleOCR

1) UPLOAD PDF FILE

Users upload invoice document files (fuel slips) via a Google Apps Script-based interface. Once uploaded, the files are automatically stored in Google Drive, while metadata such as file name, upload time, and uploader details are recorded in Google Sheets. The system then performs a validation check on the file format and required form segments. If the file format and content are valid, it proceeds to the next stage. Otherwise, an error notification pop-up is sent to the user to prompt corrections before re-uploading.

2) IMAGE PRE-PROCESSING

Once a file is deemed valid, Python processes the PDF document into images using libraries like PyPDF2 or pdf2image. OpenCV is used to refine the images before text extraction. This involves several enhancement steps:

a. Grayscale Conversion

Converting a color image to grayscale is a fundamental pre-processing step in image analysis. Color images contain three channels—red, green, and blue (RGB), which can introduce complexity in certain image processing tasks. By transforming the image into a grayscale representation, the data is simplified by preserving only the intensity information while discarding color details that may be irrelevant to the analysis. This conversion is typically performed using a standardized color transformation process, ensuring that variations in color do not affect subsequent processing stages.

b. Noise Removal

The denoising process is essential for enhancing image quality, particularly in applications such as text recognition, where noise can obscure characters and degrade readability. Images affected by noise, such as random spots or visual artifacts, may experience a loss of critical details, leading to inaccuracies in subsequent processing. A widely adopted approach for noise reduction is the Non-Local Means Denoising algorithm, which effectively suppresses small-scale noise while preserving the structural integrity of the image. This method operates by analyzing pixel similarities within a defined search window. Typical parameter values include a filter strength (h) of 30, a template window size of 7, and a search window size of 21, ensuring effective noise removal while maintaining essential image features.

c. Local Contrast Enhancement

Contrast-Limited Adaptive Histogram Equalization (CLAHE) is employed to enhance the local contrast in images, particularly in regions with low or uneven illumination. This technique partitions the image into smaller, non-overlapping blocks and applies histogram equalization individually to each segment, thereby preventing over-amplification of noise while adaptively improving contrast. To regulate contrast enhancement, a clip limit is introduced, typically set to 3.0, which restricts the amplification of high-frequency noise. Additionally, a tile grid size, commonly defined as 8×8 pixels, is used to determine the local regions for processing. This method is particularly advantageous for images affected by non-uniform lighting conditions, such as those containing blurred or barely discernible text.

d. Adaptive Thresholding

Adaptive thresholding is a technique used to convert an image into a binary (black-and-white) format, enhancing text boundaries regardless of variations in lighting across the image. Unlike global thresholding, which applies a single threshold value to the entire image, adaptive thresholding dynamically adjusts the threshold for each local region, allowing for more accurate segmentation in non-uniform lighting conditions. In this study, an adaptive Gaussian thresholding approach is utilized, with a block size of 15 and a constant subtraction factor (C) of 2. These parameters ensure optimal contrast between the text and background, improving the reliability of optical character recognition (OCR) processes.

e. Text Sharpness Enhancement

Following the thresholding process, morphological operations such as closing and dilation are applied to refine the image structure and enhance text sharpness. Closing, which consists of sequential erosion and dilation operations, is employed to fill small gaps or discontinuities within text characters. In this study, a morphological closing operation is performed using a kernel size of 2×2, ensuring the structural integrity of the text. Additionally, dilation is applied with a kernel of the same size and one iteration to expand the text regions, improving their visibility and enhancing the accuracy of optical character recognition (OCR). These morphological techniques contribute to better text continuity and readability in the processed image.

3) TEXT EXTRACTION

After pre-processing, the refined images are sent to PaddleOCR system for text recognition. PaddleOCR employs the Differentiable Binarization (DB) algorithm to detect text regions, while the Scene Text Recognition with a Single Visual Model-Lightweight Counting Network (SVTR-LCNet) model identifies and interprets text across multiple languages and document layouts. The output is raw extracted text, which will undergo further processing to identify meaningful data points.

4) INFORMATION EXTRACTION AND VALIDATION

Once the raw text has been extracted, it undergoes structured analysis using regular expressions (regex) to identify and extract key invoice details. The extracted fields typically include:

- Invoice Number
- Date
- Flight Code
- Aircraft Registration
- Departure and Destination Airport Names
- Fuel Quantity
- Vendor Name

To ensure data accuracy and consistency, the extracted information is validated using fuzzy matching techniques, which compare the retrieved text against pre-existing records in the company database. This validation process helps detect inconsistencies or anomalies, flagging potential errors for manual review and minimizing discrepancies in financial reporting.

5) STRUCTURED DATA STORAGE

Validated data is systematically stored in a structured format, such as CSV files, Google Sheets, or a cloud-based database. This structured storage enables seamless integration with external analytics tools and enterprise systems, ensuring efficient data accessibility and interoperability. The extracted data can be retrieved for further analysis or integrated with business intelligence platforms, such as Looker Studio, to facilitate advanced visualization and reporting.

The system operates through a streamlined workflow, leveraging PaddleOCR for text extraction, OpenCV for image pre-processing, Python for automation, and Google Apps Script for data management and cloud-based storage. This automated pipeline allows users to upload documents, which are then processed to extract structured data with high accuracy and efficiency.

By storing the extracted data in a centralized database, the system enhances financial data management, enabling improved decision-making, error detection, and streamlined analysis of invoice information.

B. SYSTEM INTEGRATION AND AUTOMATION

The developed system integrates various technologies to automate the invoice management workflow efficiently. This integration ensures operational reliability, accurate data extraction, and scalability for handling large volumes of documents. Main Components of System Integration:

6) GOOGLE APPS SCRIPT

Google Apps Script acts as the primary interface between users and the backend system. It manages:

- PDF file uploads.
- Storing files in Google Drive.
- Recording metadata in Google Sheets.
- Triggering subsequent automation processes like OCR and sending notifications.

7) PYTHON FOR DOCUMENT PROCESSING

Python is responsible for handling document processing tasks, including converting PDF files to images using PyPDF2 or pdf2image and pre-processing images with OpenCV to enhance OCR accuracy.

8) TEXT EXTRACTION

PaddleOCR detects and extracts text from processed images with high accuracy. The system supports complex document layouts and multiple languages, making it adaptable to diverse invoice formats.

9) DATA VALIDATION AND STORAGE

The extracted text is processed using regex and fuzzy matching techniques to validate critical information, such as invoice numbers and flight details. Once validated, the structured data is stored in:

- Google Sheets
- Cloud databases for further analysis and reporting

10) WORKFLOW AUTOMATION

To streamline operations, the system includes several automation features:

- Processed files are automatically moved to designated folders in Google Drive.
- Users receive real-time notifications about processing success or errors.

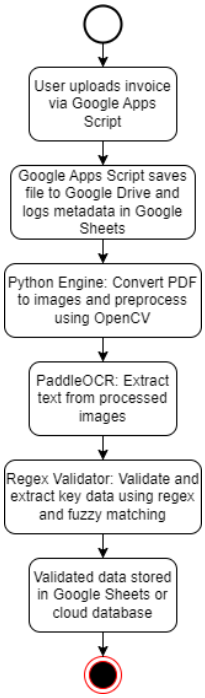


Figure 3. System Architecture Diagram for Automated Extraction of Aircraft Fuel Invoice Data

The diagram shown in Figure 3 shows the workflow of an automated data extraction system from aircraft fuel invoices. The process begins with the user uploading a PDF file, which is then processed through a preprocessing stage using OpenCV, followed by text extraction using PaddleOCR, and finally, the validated data is stored in Google Sheets or a database. Benefits of Integration and Automation:

- Flexibility: The system supports multiple document formats and varying image qualities.
- Time Efficiency: Automation reduces manual processing time from document upload to structured data storage.

- Scalability: Cloud-based infrastructure allows the system to process large volumes of invoices efficiently.
- Reliability: Seamless integration between components ensures smooth operation and minimizes errors.

C. EVALUATION AND PERFORMANCE METRICS

System evaluation was conducted to measure the effectiveness, efficiency, and accuracy of the PaddleOCR-based invoice management automation system. In this evaluation, standardization of predetermined data formats is essential to ensure better data extraction accuracy. The standardization includes:

- The ink must be clear from the original invoice.
- There is no paper or background containing letters or number characters.
- The invoice image must be visible in its entirety, without any parts being cut off.

The following are the main metrics used, and the evaluation methods applied:

11) DATA EXTRACTION ACCURACY

Accuracy is evaluated by comparing the text extracted by PaddleOCR with previously labeled ground truth data. This process is heavily influenced by the standardization of the applied format, so the extraction results are more accurate if the format is followed.

Parameters used:

- Word Error Rate (WER): Measures the percentage of correctly extracted words out of the total retrieved data.
- Character Error Rate (CER): Measures the percentage of relevant characters that were successfully extracted.

Methodology:

- The dataset consists of 100 documents and all documents meet the pre-defined format standardization.
- Each document is analyzed using PaddleOCR, and the results are evaluated against ground truth data.
- Measurements are made after ensuring that all documents meet the established format standards, so that extraction accuracy can be improved.

12) PROCESSING TIME

Processing time is evaluated to measure the efficiency of the system in handling documents of various sizes. Good format standardization will result in reduced processing time because the OCR system does not have to deal with errors or difficulties arising from low-quality documents.

Parameters measured:

- Total time from file upload to data being stored in the database in a structured format.

Evaluation process:

- The dataset contains documents with a size of 100 KB to 10 MB, all documents are uploaded after complying with the established format standards.
- Measurements are made by calculating the average processing time for each document size category.

13) SYSTEM RESILIENCE

The system's robustness is tested with documents in a variety of conditions, including documents of poorer quality. However, with good format standardization, the system can process documents more consistently and reduce the likelihood of extraction failure.

Durability is tested under the following conditions:

- Documents with complex layouts.
- Documents with low light.
- Documents with small text or high noise.

Methodology:

- PaddleOCR was tested on various data sets to evaluate whether the system still produces adequate data extraction even under adverse conditions.
- Additional testing is performed to measure the impact of document quality on accuracy, by ensuring that processed documents conform to pre-defined format standards.

14) ERROR RATE

System errors are evaluated by identifying the frequency of errors that occur during text extraction. Format standardization plays an important role in reducing errors, as poor document quality can increase error rates.

Observed parameters:

- The character is misread.
- Data does not match regex patterns.
- Data validation failed.

Evaluation Process:

- The extraction results are compared with the ground truth data to record the number of errors.
- The error frequency is calculated for each document type, considering the format standardization applied to minimize errors.

15) ERROR ANALYSIS

Each error is evaluated to determine the cause and possible solutions. Standardization of formats plays an important role in identifying and reducing errors, as OCR systems will have an easier time handling documents that meet the established quality requirements.

Methodology:

- Errors are analyzed by category: (invalid character, validation failed, or high noise).
- The system is tested with new documents that have met format standards to validate the implemented solution.

The PaddleOCR-based invoice management automation system evaluation was conducted by looking at several key metrics, which were significantly influenced by the standardization of data formats that must be met before uploading files. By following these standards, namely

ensuring that invoice ink is clear, there are no characters in the background, and the invoice image is fully visible—the data extraction process can be carried out with higher accuracy, better efficiency, and fewer errors.

III. RESULTS AND DISCUSSION

A. DATASET AND CONFIGURATION

To evaluate the performance of the automated data extraction system, we utilized a curated dataset comprising 100 real-world aircraft fuel invoice documents. These documents were sourced from actual aviation fuel transactions and reflect a wide spectrum of document quality, including high-resolution scans, blurred or faded text, noisy backgrounds, and low-light captures. This diversity ensures a rigorous evaluation of the system's robustness across realistic operational scenarios. All documents were provided in PDF format—consistent with typical industry submission practices—and varied in file size between 100 KB and 10 MB. Prior to processing, each document was screened based on the following standardization criteria:

- Clarity of Ink: The original invoices must have clear, legible ink to minimize initial text recognition errors.
- Background Cleanliness: The paper or background must be free of additional letters, numbers, or patterns that could interfere with text extraction.
- Full Visibility: The entire invoice image must be visible, with no parts cropped or obscured, to ensure all relevant data fields are accessible.

Documents failing to meet these criteria were either excluded or manually enhanced to match the standardization requirements. These rules help isolate OCR accuracy from confounding image artifacts and ensure consistency during evaluation.

For the preprocessing pipeline, we applied the following OpenCV-based techniques:

- Grayscale conversion using a standardized RGB-to-grayscale transformation.
- Non-Local Means Denoising with a filter strength (h) of 30, template window size of 7, and search window size of 21.
- CLAHE with a clip limit of 3. 0 and a tile grid size of 8×8 pixels.
- Adaptive Gaussian thresholding with a block size of 15 and a constant subtraction factor (C) of 2.
- Morphological operations (closing and dilation) using a 2×2 kernel with one iteration for dilation.

To evaluate the effectiveness of each preprocessing method, we performed a controlled ablation study, testing combinations such as grayscale only, denoising only, CLAHE only, and all combined. This allowed us to quantify each technique's individual contribution to OCR performance. Three OCR engines were benchmarked in this study:

- PaddleOCR, using the Differentiable Binarization (DB) algorithm for text detection and SVTR-LCNet for multilingual recognition.
- Tesseract OCR, configured with its standard LSTM model.
- EasyOCR, using its default transformer-based pipeline.

The system was implemented in Python and ran on a machine equipped with an Intel Core i7 processor, 16 GB RAM, and an NVIDIA GTX 1660 GPU. PDF-to-image conversion was performed using pdf2image, and each OCR result was validated against manually labeled ground truth data fields—including invoice number, date, fuel quantity, and aircraft identification. This comprehensive dataset and configuration framework formed the basis for evaluating three core metrics: data extraction accuracy, processing time, and system resilience under varied conditions.

B. RESEARCH RESULTS

1) DATA EXTRACTION ACCURACY

The system was evaluated using 100 aircraft fuel invoices that varied in quality, including high-resolution scans, blurry images, low-light exposures, and faded or noisy documents. Accuracy was assessed using two metrics: Character Error Rate (CER) and Word Error Rate (WER). PaddleOCR consistently outperformed both Tesseract and EasyOCR across all preprocessing conditions, as detailed in Table 1 and visualized in Figures 4 and 5.

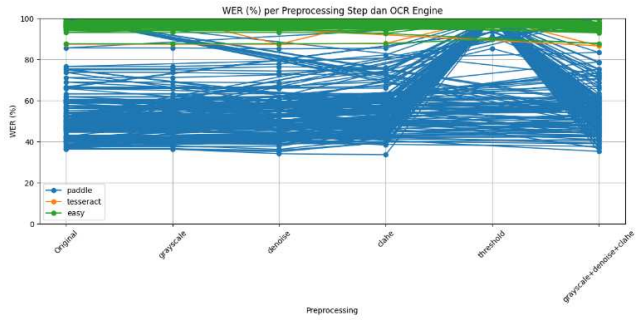
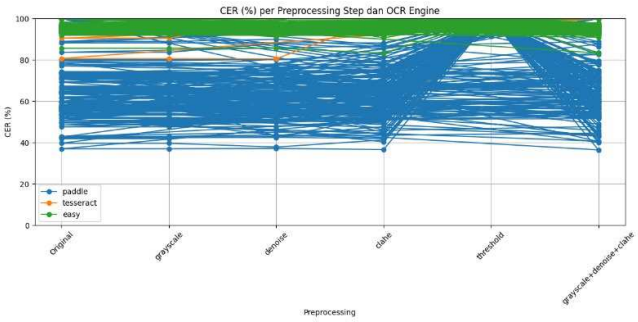


Figure 4. **World Error Rate (WER) Distribution Across 100 Aircraft Fuel**



Invoice Samples

Figure 5. **Character Error Rate (CER) Distribution Across 100 Aircraft Fuel invoice Examples**

TABLE I
OCR PERFORMANCE COMPARISON (CER AND WER) FOR PADDLEOCR, TESSERACT, AND EASYOCR ACROSS VARIOUS PREPROCESSING TECHNIQUES

OCR Engine + Preprocessing	CER (%)	WER (%)
EasyOCR	95.7265	98.1854
EasyOCR + Clahe	95.5389	97.984
EasyOCR + Denoise	95.6871	98.2225
EasyOCR + Grayscale	95.7342	98.1854
EasyOCR + Grayscale + Denoise + Clahe	95.4575	97.8651
EasyOCR + Grayscale + Denoise + Clahe + Morph	95.4575	97.8651
EasyOCR + Morph	95.7342	98.1584
EasyOCR + Threshold	99.6842	99.8035
PaddleOCR	64.3346	56.6334
PaddleOCR + Clahe	65.1385	56.4973
PaddleOCR + Denoise	64.2335	56.2613
PaddleOCR + Grayscale	64.3316	56.6204
PaddleOCR + Grayscale + Denoise + Clahe	65.0675	58.425
PaddleOCR + Grayscale + Denoise + Clahe + Morph	65.0675	58.425
PaddleOCR + Morph	64.3316	56.6204
PaddleOCR + Threshold	99.9136	99.6398
Tesseract	99.7121	99.8331
Tesseract + Clahe	100.0	99.8041
Tesseract + Denoise	99.8057	99.7056
Tesseract + Grayscale	99.7121	99.8331
Tesseract + Grayscale + Denoise + Clahe	99.9576	99.7849
Tesseract + Grayscale + Denoise + Clahe + Morph	99.9576	99.7849
Tesseract + Morph	99.7121	99.8331
Tesseract + Threshold	100.0	99.9477

Figure 4 and Figure 5 show that compared to EasyOCR and Tesseract—which showed average CER and WER values above 95% and 98%—PaddleOCR achieved much lower error rates: 64.23% CER and 56.26% WER in its raw form, with slight improvements using grayscale and denoising. This highlights PaddleOCR's superior robustness in handling complex layouts and poor-quality invoice images.

Table 1 showed that grayscale and denoising offered the most significant accuracy gains, while aggressive thresholding consistently degraded performance. These findings confirm the critical impact of document quality on OCR accuracy and reinforce PaddleOCR's scalability for noisy, real-world invoice data.

2) PROCESSING TIME

System speed is measured by the time it takes to process a document, from uploading to storing structured data in the system. An efficient average processing time (less than 60 seconds per document) shows that the system is not only effective in processing large amounts of data, but also ideal for applications that require speed, such as the aviation industry.

Figure 6 show the average processing time for each document type, which provides an idea of the system's efficiency in handling different types of documents in an aviation industry environment.

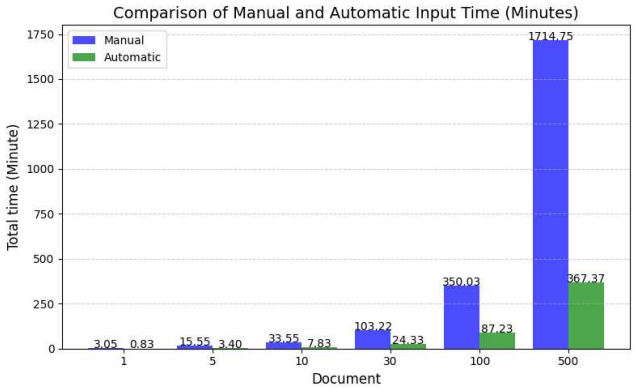


Figure 6. Comparison of Manual vs Automated Processing Time

The scalability test results show that although the processing time increases linearly with the number of documents, the system maintains stable performance even when handling large volumes.

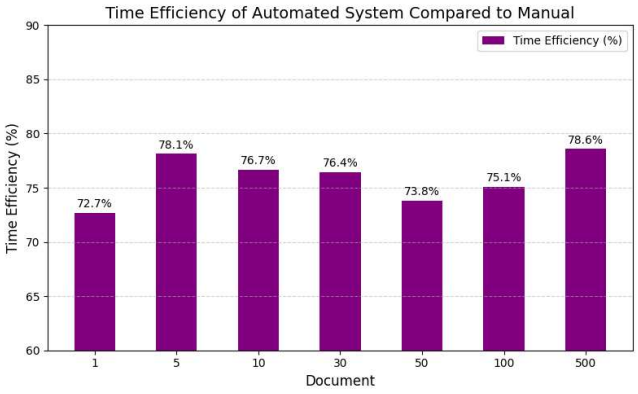


Figure 7. Comparison of Manual vs Automated Processing Time for Invoice Handling

The results of the comparison of manual and automatic processing times, as shown in Figure 7, that the system can increase time efficiency by up to 78%, while eliminating manual processes that require extra effort.

3) SYSTEM RESILIENCE

The system was tested on a variety of documents to evaluate its robustness to a variety of conditions, including low lighting, complex layouts, and text in non-standard formats. The test results showed that PaddleOCR was able to maintain an accuracy rate above 75% even in less-than-ideal conditions, such as documents with low lighting or irregular layouts. For example, on a document with low lighting in 50% of the area, the system was still able to recognize characters with 76% accuracy. However, this accuracy dropped on documents with very complex layouts or containing overlapping graphic elements.

C. DISCUSSION

1) INTERPRETATION OF DATA EXTRACTION ACCURACY RESULTS

The evaluation on 100 diverse invoice documents shows that PaddleOCR consistently delivers lower Character Error Rate (CER) and Word Error Rate (WER) than both Tesseract and EasyOCR. While Tesseract and EasyOCR suffered from over 95% CER and 98% WER in many preprocessing scenarios, PaddleOCR maintained substantially lower error rates (averaging ~64% CER and ~56% WER), especially when paired with grayscale and denoising. Ablation testing further revealed that aggressive thresholding harmed recognition accuracy across all engines, while grayscale and CLAHE offered modest improvements. These findings highlight the importance of selecting preprocessing methods carefully, and confirm that PaddleOCR offers better resilience to complex layouts, noise, and multilingual text, typical of aviation invoices.

2) TIME EFFICIENCY

The efficient average processing time (<60 seconds per document) shows that this system is suitable for industries that require fast data processing, such as the aviation industry. The speed of this system is one of the advantages in real-world applications.

3) SYSTEM ADVANTAGES

In terms of technology integration, this system combines PaddleOCR, OpenCV, and Google Apps Script to create an efficient workflow in document management. With an OCR system that is better than the previous process, the extraction process runs with even better accuracy. Other than that, automation of invoice management processes reduces manual workload and improves the accuracy of administrative data.

4) SYSTEM LIMITATIONS

This system has limitations: documents with high noise levels and small text. Documents with high noise levels affects PaddleOCR's ability to recognize characters. Other than that, very small font sizes decrease the accuracy of the system, especially on low-resolution documents. For example, to address the problem of small text, one potential solution is the application of more sophisticated deep learning technologies, such as the use of CNN-based models that are more specific to small text.

5) COMPARISON WITH PREVIOUS RESEARCH

Compared to Tesseract, this system is superior in handling complex document layouts without requiring much pre-processing adjustment. PaddleOCR also performs better than TesseractOCR and even EasyOCR in handling invoices with complex text and in multilingual text recognition, which is especially relevant for international documents.

6) RESEARCH IMPLICATIONS

The implementation of this system can reduce manual invoice processing time and improve operational efficiency in the aviation industry. This system can be applied in other sectors, such as logistics, finance, and markets, to support large-scale document management.

D. ERROR ANALYSIS

1) TYPES OF ERRORS THAT OCCUR

Character Recognition Error: Errors typically arise in documents with significant noise, low contrast, or blurred regions. For instance, in noisy documents, characters such as ‘S’ are often misinterpreted as ‘5’, especially by EasyOCR and Tesseract. PaddleOCR exhibits more robustness but still occasionally fails under extreme conditions.

- Over-Thresholding Artifacts: Aggressive thresholding (especially global or Gaussian) introduces artifacts that disconnect character strokes or over-simplify edges, reducing text recognition accuracy. This error was prevalent across all OCR engines when thresholding was used without grayscale or denoising.
- Malformed Data Structures: When document layouts deviate from the expected invoice format (e.g., flight code placed in a non-standard position or use of unusual separators), regex patterns fail to extract data accurately.
- Validation Errors: Data mismatch issues arise when OCR returns partial or corrupted values (e.g., missing prefixes in flight codes) that don’t align with company records. Fuzzy matching can mitigate some, but not all, of these discrepancies.

2) PROPOSED SOLUTION

- Pre-processing Refinement: Avoid aggressive thresholding as default. Favor grayscale combined with denoising and CLAHE, which statistically provided the best balance in accuracy. Consider adaptive preprocessing selection based on document condition (e.g., brightness histogram).
- OCR Engine Fine-tuning: Develop custom PaddleOCR model trained with domain-specific (aviation invoice) samples, including blurred, low-light, and rotated examples. This will enhance robustness against real-world data irregularities.
- Regex Flexibility and Fallbacks: Expand regex coverage to include variant invoice patterns and implement fallback heuristics (e.g., proximity-based token detection) when standard patterns fail.
- Confidence-based Filtering: Integrate character-level confidence scores (available in PaddleOCR) to reject low-confidence fields or trigger human review, improving validation accuracy and reducing false entries.

IV. CONCLUSION

This research successfully developed an automated invoice document management system using PaddleOCR, enhancing operational efficiency in the aviation industry.
Key findings include:

1) TIME EFFICIENCY AND ACCURACY

The system reduces manual processing time, achieving an average processing speed of under 60 seconds per document with high accuracy for good-quality inputs.

2) SYSTEM PERFORMANCE

PaddleOCR performs well on standard documents but faces challenges of high noise or small text, which can be improved through deep learning-based pre-processing.

3) SCALABILITY AND FUTURE DEVELOPMENT

The system processes up to 640 documents efficiently, with potential enhancements including improved OCR algorithms, and ERP integration.

4) ECONOMIC IMPACT

Automation optimizes operational costs by identifying transaction discrepancy, improving data accuracy, and enhancing workforce efficiency for strategic processes.

This study demonstrates that PaddleOCR-based automation significantly improves document processing, with promising opportunities for further optimization and cross-industry applications.

V. RECOMMENDATION

To optimize the PaddleOCR-based invoice automation system, several focused improvements are recommended based on the expanded dataset and comparative evaluation:

- Enhancing Accuracy Avoid aggressive thresholding, as it reduces performance. Instead, prioritize grayscale conversion, denoising (preferably deep learning-based), and CLAHE. Training PaddleOCR with aviation-specific data can boost recognition of flight and fuel details.
- Improving Speed Maintain <60s per document by adding queue-based processing and leveraging cloud autoscaling to handle growing volumes efficiently.
- Strengthening System Integration Integrate with ERP systems (e.g., SAP, Oracle) via secure APIs. Expand support for various file formats and multilingual OCR for broader applicability.
- Handling Handwritten Input Add support for handwriting OCR (e.g., Google Cloud Vision, Microsoft OCR) to process documents with mixed printed and handwritten text.
- Smarter Data Validation Combine regex with AI-based validation and fuzzy matching to handle format variations and reduce mismatches, especially for vendor names and airport codes.

AUTHORS CONTRIBUTION

Reinaldy Hutapea: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Software, Visualization, Writing – Original Draft Preparation;
Vanessa Harwanto: Conceptualization, Methodology, Project Administration, Resources, Supervision, Validation, Visualization;
Samuel Situmeang: Project Administration, Supervision, Validation, Writing – Review & Editing;

COPYRIGHT

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

REFERENCES

[1] M. P. Mantje, P. Rambe, and T. M. Ndofirepi, "Effects of knowledge management on firm competitiveness: The mediation of operational efficiency," *SA Journal of Information Management*, vol. 25, no. 1, pp. 1–10, 2023, doi: 10.4102/sajim.v25i1.1549.

[2] I. Farida and D. Setiawan, "Business Strategies and Competitive Advantage: The Role of Performance and Innovation," *Journal of Open Innovation: Technology, Market, and Complexity*, vol. 8, no. 3, p. 163, 2022, doi: 10.3390/joitmc8030163.

[3] X. Teng, Z. Wu, and F. Yang, "Research on the Relationship between Digital Transformation and Performance of SMEs," *Sustainability (Switzerland)*, vol. 14, no. 10, pp. 1–17, 2022, doi: 10.3390/su14106012.

[4] D. Baviskar, S. Ahirrao, V. Potdar, and K. Kotecha, "Efficient Automated Processing of the Unstructured Documents Using Artificial Intelligence: A Systematic Literature Review and Future Directions," *IEEE Access*, vol. 9, pp. 72894–72936, 2021, doi: 10.1109/ACCESS.2021.3072900.

[5] N. L. Rane, S. P. Choudhary, and J. Rane, "Artificial Intelligence-Driven Corporate Finance: Enhancing Efficiency and Decision-Making Through Machine Learning, Natural Language Processing, and Robotic Process Automation in Corporate Governance and Sustainability," *Studies in Economics and Business Relations*, vol. 5, no. 2, pp. 1–22, Jun. 2024, doi: 10.48185/sebr.v5i2.1050.

[6] E. Karakilic, E. Gunaltili, S. Ekici, A. Dalkiran, O. Balli, and T. H. Karakoc, "A Comparative Study between Paper and Paperless Aircraft Maintenance: A Case Study," *Sustainability (Switzerland)* , vol. 15, no. 20, 2023, doi: 10.3390/su152015150.

[7] IBISWorld, "Market Size of The Airline Industry Worldwide from 2018 to 2023 (in Billion U.S. Dollars)." Accessed: Mar. 08, 2025. [Online]. Available: https://www.statista.com/statistics/1110342/market-size-airline-industry-worldwide

[8] T3 Technology Hub, "Market Penetration of Document Processing Software Worldwide in 2021 to 2024." Accessed: Mar. 08, 2025. [Online]. Available: https://www.statista.com/statistics/1331284/document-processing-software-market-share

[9] H. Simanjuntak, E. Panjaitan, S. Siregar, U. Manalu, S. Situmeang, and A. Barus, "Generating New Ulos Motif with Generative AI Method in Digital Tenun Nusantara (DiTenun) Platform," *International Journal of Advanced Computer Science and Applications*, vol. 15, no. 7, 2024, doi: 10.14569/IJACSA.2024.01507109.

[10] N. D. Hutahaean, I. J. Siregar, A. S. Lumbantobing, P. R. Togatorop, H. T. A. Simanjuntak, S. I. G. Situmeang, and A. B. Lumbantobing, "Data Integration Automation from Heterogeneous Data Sources for Smart Farming Data Lake," 2025, p. 050006. doi: 10.1063/5.0241665.

[11] Viavi Solutions, "What Are The Top Two Benefits of Using Automation? [Graph]." Accessed: Mar. 08, 2025. [Online]. Available: https://www.statista.com/statistics/1298858/benefits-of-using-automation-worldwide

[12] G. A. Cutting and A.-F. Cutting-Decelle, "Intelligent Document Processing -- Methods and Tools in The Real World," pp. 1–28, Dec. 2021, [Online]. Available: http://arxiv.org/abs/2112.14070

[13] R. Darmawan, A. Nasuha, L. Zaman, and H. Armanto, "Sistem Manajemen Kartu Nama dengan OCR dan Ekstraksi Informasi Otomatis," *Journal of Intelligent System and Computation*, vol. 3, no. 2, pp. 61–72, Oct. 2021, doi: 10.52985/insyst.v3i2.194.

[14] C. Irimia, F. Harbuzariu, I. Hazi, and A. Iftene, "Official Document Identification and Data Extraction using Templates and OCR," *Procedia Computer Science*, vol. 207, pp. 1571–1580, 2022, doi: 10.1016/j.procs.2022.09.214.

[15] J. Krishna Manipatruni, R. Gnana Sree, R. Padakanti, S. Naroju, B. Kumar Depuru, R. Associate, and C. Author, "Leveraging Artificial Intelligence for Simplified Invoice Automation: Paddle OCR-Based Text Extraction from Invoices," *International Journal of Innovative Science and Research Technology*, vol. 8, no. 9, 2023.

[16] S. V. Mahadevkar, S. Patil, K. Kotecha, L. W. Soong, and T. Choudhury, "Exploring AI-driven approaches for unstructured document analysis and future horizons," *Journal of Big Data*, vol. 11, no. 1, 2024, doi: 10.1186/s40537-024-00948-z.

[17] S. Carta, A. Giuliani, L. Piano, and S. G. Tiddia, "An End-to-End OCR-Free Solution For Identity Document Information Extraction," *Procedia Computer Science*, vol. 246, pp. 453–462, 2024, doi: 10.1016/j.procs.2024.09.425.

[18] T. Hegghammer, "OCR with Tesseract, Amazon Textract, and Google Document AI: a benchmarking experiment," *Journal of Computational Social Science*, vol. 5, no. 1, pp. 861–882, 2022, doi: 10.1007/s42001-021-00149-1.

[19] M. Modi, A. Shah, D. Kothadiya, and M. Rahevar, "Optical Character Recognition: Comparative Analysis of Tesseract and Textract on Diverse Datasets," in *2024 3rd International Conference on Automation, Computing and Renewable Systems (ICACRS)*, IEEE, Dec. 2024, pp. 913–919. doi: 10.1109/ICACRS62842.2024.10841500.

[20] S. Kumar, N. K. Sharma, M. Sharma, and N. Agrawal, "Text Extraction from Images Using Tesseract," in *Deep Learning Techniques for Automation and Industrial Applications*, Wiley, 2024, pp. 1–18. doi: 10.1002/9781394234271.ch1.

[21] Y. Li, J. Poon, and U. Roehm, "Synergizing Optical Character Recognition: A Comparative Analysis and Integration of Tesseract, Keras, Paddle, and Azure OCR," 2024.

[22] E. T. P. Zebua and P. Rosyani, "Perancangan Deteksi Objek Kendaraan Bermotor Berbasis OpenCV Python menggunakan Metode HOG-SVM untuk Analisis Lalu Lintas Cerdas," *Jurnal AI dan SPK : Jurnal Artificial Inteligent dan Sistem Penunjang Keputusan*, vol. 2, no. 1, pp. 16–26, 2024.

[23] M. Haris, M. G. Suryanata, and M. Yetri, "Implementasi OCR Menggunakan Algoritma Template Matching Correlation pada Pengarsipan e-KTP," *J-SISKO TECH (Jurnal Teknologi Sistem Informasi dan Sistem Komputer TGD)*, vol. 6, no. 2, p. 281, 2023, doi: 10.53513/jsk.v6i2.8134.