Matrik: Jurnal Manajemen, Teknik Informatika, dan Rekayasa Komputer
Vol. 24, No. 2, March 2025, pp. 195∼206
ISSN: 2476-9843, accredited by Kemenristekdikti, Decree No: 200/M/KPT/2020
DOI: 10.30812/matrik.v24i2.4367

Analyzing the Application of Optical Character Recognition: A Case Study in International Standard Book Number Detection

Imam Fahrur Rozi, Ahmadi Yuli Ananta, Endah Septa Sintiya, Astrifidha Rahma Amalia, Yuri Ariyanto, Arin Kistia Nugraeni
Politeknik Negeri Malang, Malang, Indonesia

Article Info

Article history:

Received August 09, 2024
Revised November 03, 2024
Accepted January 07, 2025

Keywords:

Book Number Detection;
International Standard;
Optical Character Recognition.

ABSTRACT

In the era of advanced education, assessing lecturer performance is crucial to maintaining educational quality. One aspect of this assessment involves evaluating the textbooks authored by lecturers. This study addresses the problem of efficiently detecting International Standard Book Numbers (ISBNs) within these textbooks using optical character recognition (OCR) as a potential solution. The objective is to determine the effectiveness of OCR, specifically the Tesseract platform, in facilitating ISBN detection to support lecturer performance assessments. The research method involves automated data collection and ISBN detection using Tesseract OCR on various sections of textbooks, including covers, tables of contents, and identity pages, across different file formats (JPG and PDF) and orientations. The study evaluates OCR performance concerning image quality, rotation, and file type. Results of this study indicate that Tesseract performs effectively on high-quality, low-noise JPG images, achieving an F1 score of 0.97 for JPG and 0.99 for PDF files. However, its performance decreases with rotated images and certain PDF conditions, highlighting specific limitations of OCR in ISBN detection. These findings suggest that OCR can be a valuable tool in enhancing lecturer performance assessments through efficient ISBN detection in textbooks.

Corresponding Author:

Imam Fahrur Rozi, +6285233139738,
Faculty of Information Technology,
Politeknik Negeri Malang, Malang, Indonesia,
Email: imam.rozi@polinema.ac.id.

How to Cite:
I. Rozi, A. Ananta, E. Sintiya, A. Amalia, Y. Ariyanto, and A. Nugraeni, "Analyzing the Application of Optical Character Recognition: A Case Study in International Standard Book Number Detection", MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer, vol. 24, no. 2, pp. 195-206, Feb. 2025.

1. INTRODUCTION

Lecturers play a crucial role in enhancing the quality of human resources and advancing the nation's intellectual life. Every lecturer is required to improve themselves and maintain high performance continuously [1]. Lecturer performance refers to the ability to perform tasks or responsibilities assigned to them in completing their work [2, 3]. For institutions, the overall performance of lecturers significantly impacts various areas, including accreditation and the quality of education delivery [2]. Furthermore, accreditation is crucial for educational institutions, serving as a key attraction for prospective students [4] and as a competitive advantage to enhance overall competitiveness. Lecturer performance is one of the key factors in the institutional accreditation assessment process. Higher education institutions need to foster a high level of work motivation and a conducive work environment to enhance lecturer performance [2]. However, evaluating lecturer performance, particularly in the publication of International Standard Book Numbers (ISBNs)-registered textbooks, remains challenging and time-consuming when reliant on manual verification, often delaying assessments that impact institutional evaluations. One aspect that can facilitate this evaluation is the ability to detect ISBNs in textbooks authored by lecturers, streamlining the assessment process. This study proposes using optical character recognition (OCR) technology, specifically through the Tesseract platform, to automate ISBN detection in lecturer-authored textbooks. By integrating OCR technology, this approach aims to replace manual processes and improve the efficiency and accuracy of lecturer performance assessments. This research aims to evaluate the effectiveness of OCR technology in detecting ISBNs across various image conditions and file formats. Through this, the study seeks to enhance the ISBN recognition process, providing institutions with a reliable tool to support lecturer performance evaluations related to textbook publications, ultimately leading to more timely and accurate assessments.

Previous studies have indicated that OCR is used for text recognition in document images, whether it be printed text, handwriting, or text in background images [5]. OCR is used to convert text in images into text that a computer can process [6]. This is essential in the document digitization, where the text in document images needs to be converted into text that can be indexed, searched, and analyzed computationally. Using OCR, an effective approach can be implemented to recognize text in document images and improve text recognition accuracy [7], [8]. While OCR technology has been successfully applied in numerous document recognition contexts, this study focuses on a specialized application: ISBN detection in educational texts authored by lecturers. Other studies have discussed the implementation of OCR systems using Tesseract and classification techniques to automate the extraction of information from paper documents and reduce the need for manual data entry [8]. Tesseract is a library that provides an OCR engine with support for Unicode and the capability to recognize over 100 languages out-of-the-box [9]. Tesseract can also be trained to recognize other characters and languages [10]. Tesseract is one of the OCR engines that has seen significant performance improvements thanks to the integration of machine-learning techniques [11]. Tesseract's capabilities extend to processing images that are rotated, tilted, or skewed, and it supports multiple languages and scripts, making it suitable for document scanning, archiving, and indexing [12]. In practical applications, such as invoice processing, Tesseract can automate the extraction of text from scanned invoices, significantly improving efficiency and accuracy [13]. Additionally, Tesseract has been adapted for specialized tasks, such as recognizing the Tifinagh script, where a comprehensive dataset was created to train the OCR engine, achieving high accuracy and precision [14]. Although Tesseract has been widely used for various OCR applications, including specialized tasks, studies applying Tesseract specifically to ISBN detection for lecturer performance assessment remain limited. This research addresses this gap by implementing Tesseract OCR to automate ISBN verification in academic settings, offering a novel approach to enhancing lecturer performance evaluations.

This study evaluates lecturer performance in producing ISBN-registered textbooks by applying OCR technology to automate ISBN detection. Previous research has not resolved some gaps, namely the limited application of OCR technology, specifically Tesseract, for ISBN detection in educational texts authored by lecturers. By addressing the gap in applying OCR, specifically tesseract, to ISBN detection for performance assessment, the research aims to enhance the efficiency and accuracy of lecturer evaluations. Traditional methods of evaluating lecturer performance in producing textbooks are time-consuming and may delay institutional assessments. This process can be automated with OCR, allowing quicker and more reliable evaluations. The findings are expected to have broader implications for educational institutions by supporting faster, more accurate assessments that directly enhance institutional quality measures and accreditation processes. The approach taken in this evaluation encompasses various critical aspects, such as book covers, tables of contents, identity pages, and ISBN-based verification. Automated data collection and ISBN detection are performed using OCR technology with the open-source Tesseract library. Tesseract was chosen as the tool to implement OCR in this research due to its high accuracy and widespread use, making it suitable for practical applications [15]. The difference between this research and the previous one is that this study implements Tesseract OCR to automate ISBN detection in lecturer-authored textbooks, addressing the identified gap in applying OCR technology for performance assessment. This study contributes by designing and implementing an innovative lecturer performance evaluation system, focusing on creating ISBN-registered textbooks, using Tesseract OCR for ISBN detection across various image conditions and file formats.

With the application of OCR technology, the evaluation of lecturer performance related to ISBN-registered book publications is expected to be conducted more efficiently and produce more organized data. OCR technology can read characters from physical text or images, allowing this process to occur more quickly and efficiently. This research contributes to the development of science by designing and implementing an innovative lecturer performance evaluation system that utilizes Tesseract OCR for ISBN detection across various image conditions and file formats. By addressing the gap in applying OCR technology specifically for performance assessment, this study enhances the efficiency and accuracy of evaluating lecturer-authored textbooks, thereby supporting educational institutions' accreditation processes and quality measures. Additionally, the results of this study are expected to provide valuable insights into the factors contributing to OCR accuracy and positively impact the accuracy of lecturer performance evaluations. Furthermore, the ISBN detection system can be an initial step in developing technology to support the lecturer's performance evaluation process.

2. RESEARCH METHOD

2.1. Research Flow

A systematic methodology was employed to manage this study effectively, as shown in Figure 1. The approach adopted in this investigation involves a multi-step method formulated to accurately detect ISBNs displayed on textbook covers by employing OCR. Initially, photographs of textbook covers are taken at a high resolution to ensure clear visibility of the ISBN area. Following this, the images are formatted into the appropriate file formats, like JPG and PDF, to standardize the input for the OCR process. These two file types were selected due to their frequent use by users when uploading documents. The central component of the approach utilizes an OCR module based on Tesseract, which extracts textual data focusing on identifying ISBNs usually situated in specific areas on the covers. To assess the accuracy of the OCR module, a ground truth dataset is established by manually marking the correct ISBNs on a selection of images. The OCR system's performance is then tested using established performance metrics, including Precision, Recall, and F1-Score, which collectively assess the system's precision in recognizing ISBNs. This systematic methodology is intended to showcase the feasibility of Tesseract-based OCR for ISBN identification, potentially enhancing cataloging and organization in educational environments.

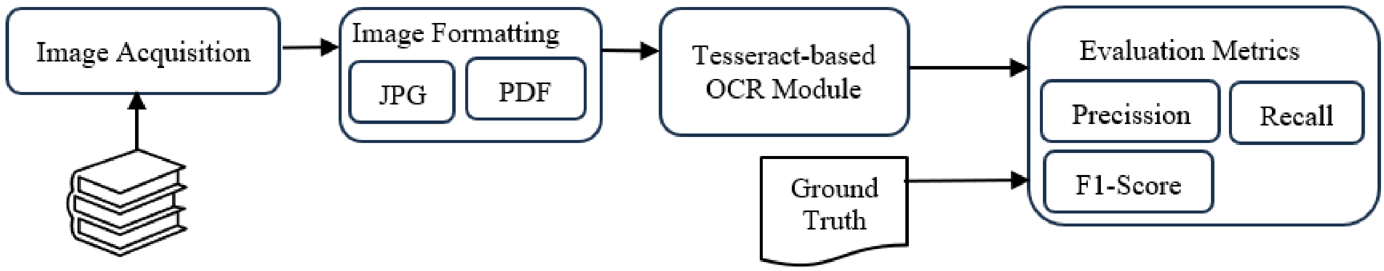

Figure 1. Research Flow

The following flowchart illustrates the OCR processing stages using Tesseract for ISBN detection in textbooks to clarify the research process. The flowchart in Figure 2 outlines five main stages: Image Acquisition, where textbook images are captured; Image Formatting into JPG and PDF formats; OCR Processing with Tesseract to extract ISBN text; Ground Truth Comparison to verify results against manually validated data; and Evaluation Metrics including Precision, Recall, and F1 Score to assess accuracy. Each step ensures the OCR system's effectiveness under various conditions.

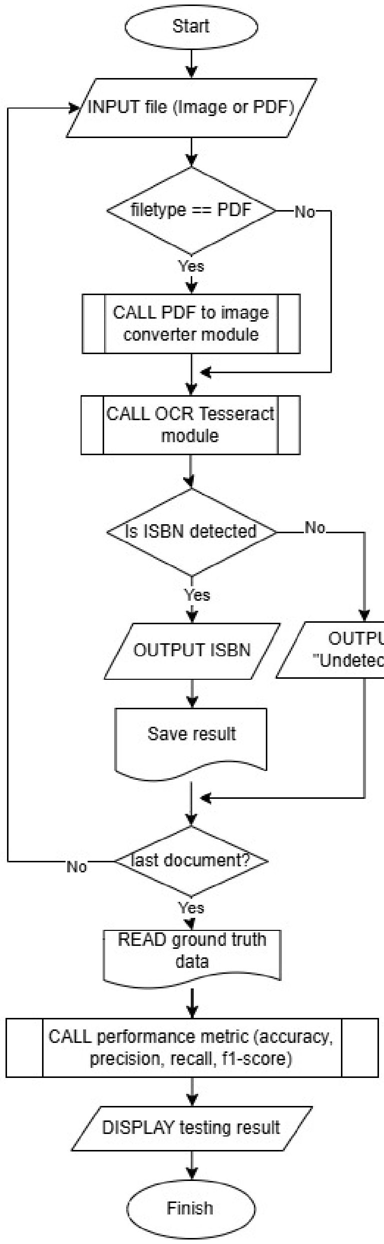

Figure 2. Flowchart of the OCR Process

2.2. Data

This research uses primary data collected directly by the researchers through field observations. The data consists of images from book covers and identity pages in various file formats commonly used by users. The data were collected by scanning several sample books' cover pages, tables of contents, and identity pages. To evaluate the accuracy of ISBN detection using Tesseract, the data were collected under several varied conditions, including (1) scanned images, (2) images rotated at 10, 30, 60, 90, 180, and 270 degrees, (3) images with high noise levels and poor lighting, and (4) images with low noise levels and good lighting. These image conditions were deliberately chosen to represent the various situations users might encounter when capturing images of book covers and identity pages.

2.3. Tesseract for OCR

This section details the approach employed in this research for implementing OCR using Tesseract. OCR is a technology that convert scanned images or printed text into text that machines can read [11]. The application of OCR technology involves examining and comprehending a variety of documents or images and facilitating their conversion into data that is amenable to analysis, modification, and retrieval. This technology has found utility in the process of digitizing handwritten medieval manuscripts, transforming typed documents into a digital format, and safeguarding historical records, legal papers, and educational resources [5]. Tesseract is one of the most popular open-source OCR engines that is ready for machine-learning integration and supports Unicode [9]. Various programming platforms widely support it, so it has gained extensive popularity in the field of OCR. It has undergone notable enhancements, particularly through the incorporation of various machine-learning methods [11, 16]. Users can further train Tesseract to recognize new languages or improve recognition accuracy for specific languages or fonts. The system developed in this study uses Tessearct as the OCR engine and is integrated with the web application system.

2.4. Evaluation Metrics

To evaluate the performance of the OCR system, we employed key metrics including recall, precision, and F1-score. These metrics provide a comprehensive assessment of the system's accuracy in identifying and recognizing text, allowing us to quantify its effectiveness and identify areas for improvement. Recall is a metric that quantifies the accuracy in recognizing characters, representing the proportion of correctly identified characters to the total number of characters requiring identification. It is determined by dividing the total number of accurately identified words by the model by the total number of words as depicted in Equation 1 [8]. By calculating recall value, we can measure the proportion of ISBN characters correctly identified among the total ISBN characters that should be detected.

$$Recall = \frac{TP}{TP + FP} \tag{1}$$

TP or True Positive is the correct OCR result. OCR results are considered correct if the OCR output matches the actual text. FP or False Positive is the incorrect OCR result. OCR results are believed to be false if the OCR output does not match the actual text. FN or False Negative is the missed OCR result. OCR results are considered missed if the OCR output is not present in the actual text [16, 17]. We also utilized precision to evaluate character recognition accuracy by determining the proportion of accurately identified characters among all characters recognized by the OCR system. This metric, precision, is computed by dividing the total number of correctly identified words by the model by the overall number of words detected. The formula for precision, as shown in Equation 2, involves a comparison between the correctly identified words and the total detected words, providing valuable insights into the performance of the OCR model [8] Precision tells us how accurately the OCR system identifies ISBN characters among all characters it has detected. It specifically answers the question: Of all the characters the OCR marked as ISBNs, how many are truly ISBNs?

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

Another evaluation metric that is utilized in this research is the F1-Score. It combines precision and recall into a single value, evaluating how well the OCR system recognizes characters. The aggregation value considers both precision, which looks at how accurately the system identifies relevant characters, and recall, which measures how well the system captures all the relevant characters. The F1 score is calculated using the Equation 3 [8], balancing precision and recall to provide a complete assessment of the OCR's performance. The F1 score provides a balanced measure of the OCR system's accuracy by combining precision and recall. It is particularly useful when we want a single metric to reflect both the completeness and correctness of ISBN detection.

$$F1 - Score = \frac{2 * (Precision * Recall)}{Precision + Recall} \tag{3}$$

3. RESULT AND ANALYSIS

The system for detecting ISBNs is constructed by utilizing the Tesseract library in conjunction with the programming language Python. Various rigorous tests were performed on different image types to evaluate the system's performance. These included scanned images, photographs captured using a smartphone camera that exhibited low resolution, significant levels of noise, and inadequate lighting conditions. Additionally, the system was evaluated using images deliberately rotated at different angles, including 10, 30, 60, 90, 180, and 270 degrees, to investigate its robustness under varying orientations. The system's capability was also assessed on images with reduced noise levels and optimal lighting conditions to ascertain its performance under more favorable settings. The set of images utilized for conducting these comprehensive tests can be observed in Figure 3, providing a visual representation of the diverse range of scenarios under which the ISBN detection system was rigorously evaluated.

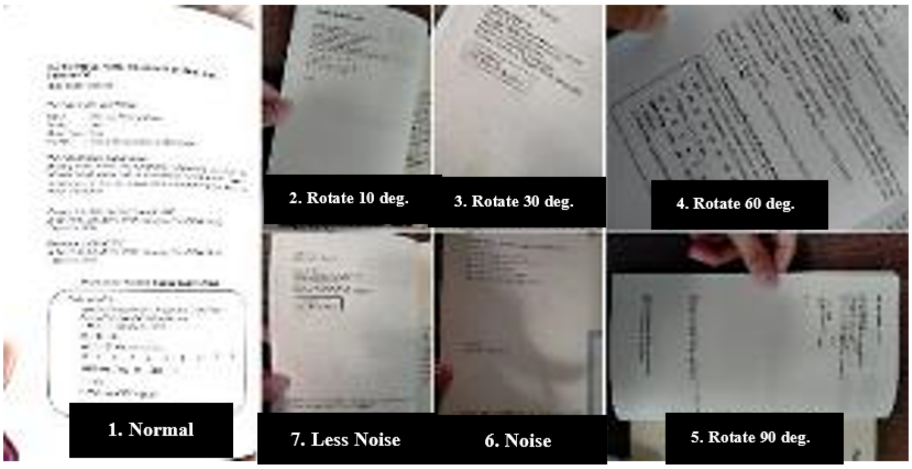

Figure 3. Test Image Samples

3.1. OCR Performance on PDF File Formats

This testing scenario is carried out to evaluate the performance of the resulting OCR system if it is applied to PDF files. Throughout this evaluation, a series of tests will be conducted, and the document's front cover, table of contents, and identification page will be obtained under different conditions to encompass and consider the potential discrepancies that may arise in various image-capturing situations. Examining the system's performance in managing PDF documents comprehensively is essential for understanding its potential and drawbacks in real-world use cases, which can drive advancements in OCR technology. Testing was carried out on 10 images in PDF file format taken using Adobe Scan, a scanner application on a smartphone. Each TP, FN, and FP value is calculated from the ISBN character. Table 1 depicts the result from the various object positions, including normal position, 10^o, 3^o, 60^o, and 90^o rotation. The systematic presentation of these findings in tabular format facilitates a clear understanding of the experimental results and allows comparisons across different scenarios.

Table 1. Evaluation metric for PDF files in normal position, 10^o, 30^o, 60^o, and 90^o rotation

No	Data	Normal Position			Rotate 10^o			Rotate 30^o			Rotate 60^o			Rotate 90^o		
		TP	FN	FP	TP	FN	FP	TP	FN	FP	TP	FN	FP	TP	FN	FP
1	Data 1	15	2	2	0	17	0	0	17	0	0	17	0	17	0	1
2	Data 2	17	0	0	0	13	0	0	13	0	0	13	0	13	0	0
3	Data 3	17	0	0	0	17	0	0	17	0	0	17	0	17	0	0
4	Data 4	17	0	0	0	17	0	0	17	0	0	17	0	17	0	0
5	Data 5	13	0	0	0	17	0	0	17	0	0	17	0	17	0	0
6	Data 6	17	0	0	0	17	0	0	17	0	0	17	0	17	0	0
7	Data 7	17	0	0	0	13	0	0	13	0	0	13	0	13	0	0
8	Data 8	13	1	1	0	17	0	0	17	0	0	17	0	16	1	1
9	Data 9	12	1	1	0	17	0	0	17	0	0	17	0	17	0	0
10	Data 10	17	0	0	0	17	0	0	17	0	0	17	0	17	0	0
Total		155	4	4	0	162	0	0	162	0	0	162	0	161	1	2

In the normal position, the OCR system correctly detected 155 ISBN characters. In this position, the system could not identify correctly 4 ISBN characters (FN) and four characters where the system incorrectly identified something as an ISBN character (FP). When tested with data rotated at 10, 30, and 60 degrees, the system failed to detect any ISBN characters, resulting in an FN value of 162. However, the system performed better with data rotated at 90 degrees, correctly identifying 161 ISBN characters, missing 1 ISBN character, and incorrectly identifying two non-ISBN characters as ISBNs. Table 2 depicts the results obtained from various scenarios, all of which stay for PDF data. The data in Table 2 represents the performance metrics associated with data corresponding to rotation angles of 180 and 270 degrees and data with different noise levels. Specifically, the table shows the value of test metrics for data exposed to high and low noise levels. Data characterized by high noise levels were included in the analysis, which was carried out under less-than-optimal lighting conditions. The test metric values presented in Table 2 provide comprehensive insight into the impact of rotational position and noise intensity on data performance. In the rotating position 180 and 270 degrees, the system could not correctly detect any ISBN character from the data provided. When applied to noisy data, the system could only identify correctly 22 ISBN characters. There were 144 characters identified incorrectly as ISBN characters and 11 non-ISBN characters incorrectly identified as ISBNs. The best performance could be achieved when the system is applied to less noisy data. In this case, there were 161 ISBN characters classified correctly, only 1 ISBN character identified incorrectly, and two non-ISBN characters identified incorrectly as ISBN characters.

Table 2. Evaluation metric for PDF files in 180^o, 270^o rotation, noisy and less noise image

No	Data	Rotate 180^o			Rotate 270^o			Noisy image			Less noise image		
		TP	FN	FP	TP	FN	FP	TP	FN	FP	TP	FN	FP
1	Data 1	0	17	0	0	17	0	0	17	0	17	0	1
2	Data 2	0	13	0	0	13	0	0	17	0	13	0	0
3	Data 3	0	17	0	0	17	0	0	17	0	17	0	0
4	Data 4	0	17	0	0	17	0	0	17	0	17	0	0
5	Data 5	0	17	0	0	17	0	5	12	10	17	0	0
6	Data 6	0	17	0	0	17	0	0	13	0	17	0	0
7	Data 7	0	13	0	0	13	0	0	17	0	13	0	0
8	Data 8	0	17	0	0	17	0	17	0	1	16	1	1
9	Data 9	0	17	0	0	17	0	0	17	0	17	0	0
10	Data 10	0	17	0	0	17	0	0	17	0	17	0	0
Total		0	162	0	0	162	0	22	144	11	161	1	2

3.2. OCR Performance on JPG File Formats

We conducted a different testing scenario to evaluate how well the system performs with JPG image files. This scenario involved applying a series of tests similar to those in the previous section but with a different file type. While the previous tests focused on the system's performance with PDF documents, this testing focused on JPG image files. We tested 10 JPG images taken with a smartphone camera. TP, FN, and FP values are still calculated based on the ISBN characters. Table 3 shows the metric value from various object positions, including the normal position. In the normal position, results similar to the testing results of the PDF files were achieved. It correctly detected 155 ISBN (TP) characters, but it could not identify correctly 4 ISBN characters (FN) and four characters where the system incorrectly identified something as an ISBN character (FP). If applied to data rotated at 10, 30, and 60 degrees, the system failed to detect any ISBN characters correctly, resulting in an FN value of 162. The better results were achieved when the system supplied data rotated at 90 degrees. It correctly identified 148 ISBN characters (TP), missing 18 ISBN characters (FN), and incorrectly identified one non-ISBN character as ISBN (FP).

Table 3. Evaluation metric for JPG files in normal position, 10^o, 30^o, 60^o and 90^o rotation

No	Data	Normal Position			Rotate 10^o			Rotate 30^o			Rotate 60^o			Rotate 90^o		
		TP	FN	FP	TP	FN	FP	TP	FN	FP	TP	FN	FP	TP	FN	FP
1	Data 1	15	2	2	0	17	0	0	17	0	0	17	0	0	17	0
2	Data 2	17	0	0	0	13	0	0	13	0	0	13	0	13	0	0
3	Data 3	17	0	0	0	17	0	0	17	0	0	17	0	17	0	0
4	Data 4	17	0	0	0	17	0	0	17	0	0	17	0	17	0	0
5	Data 5	13	0	0	0	17	0	0	17	0	0	17	0	17	0	0
6	Data 6	17	0	0	0	17	0	0	17	0	0	17	0	17	0	0
7	Data 7	17	0	0	0	13	0	0	13	0	0	13	0	13	0	0
8	Data 8	13	1	1	0	17	0	0	17	0	0	17	0	16	1	1
9	Data 9	12	1	1	0	17	0	0	17	0	0	17	0	17	0	0
10	Data 10	17	0	0	0	17	0	0	17	0	0	17	0	17	0	0
Total		155	4	4	0	162	0	0	162	0	0	162	0	148	18	1

Essentially, the scenario applied in Table 4 is the same as in Table 2 but using JPG file format. From Table 4, similar results to those in Table 2 were obtained in the testing scenarios with noisy data and less noisy data. For noisy data, both scenarios showed poor detection results, indicated by a high FN value of 132 for the JPG format. On the other hand, for less noisy data, good results were achieved, as indicated by a high TP value of 145 for JPG format data. A different outcome was observed for data with 180^o and 270^o rotations. In PDF format, the system did not perform well; however, with JPG data, the system performed well. This is evidenced by the high TP value of 144 for both 180^o and 270^o rotations. Although some ISBN characters were still undetected (FN), and some non-ISBN characters were incorrectly identified as ISBNs (FP), these errors were minimal.

Table 4. Evaluation metric for JPG files in 180^o, 270^o rotation, noisy and less noise image

No	Data	Rotate 180^o			Rotate 270^o			Noisy image			Less noise image		
		TP	FN	FP	TP	FN	FP	TP	FN	FP	TP	FN	FP
1	Data 1	0	17	0	0	17	0	0	17	0	0	17	0
2	Data 2	13	0	0	13	0	0	0	17	0	13	0	0
3	Data 3	17	0	0	17	0	0	0	17	0	17	0	0
4	Data 4	17	0	0	17	0	0	0	17	0	17	0	0
5	Data 5	17	0	0	17	0	0	17	0	0	17	0	0
6	Data 6	17	0	0	17	0	0	0	13	0	17	0	0
7	Data 7	13	0	0	13	0	0	0	17	0	13	0	0
8	Data 8	16	1	1	16	1	1	17	0	0	17	1	1
9	Data 9	17	0	0	17	0	0	0	17	0	17	0	0
10	Data 10	17	0	0	17	0	0	0	17	0	17	0	0
Total		144	18	1	144	18	1	34	132	0	145	18	1

3.3. Discussions

An experiment was carried out on 90 images in PDF and JPG formats, encompassing 9 distinct scenarios. These scenarios included images in standard or normal orientation, images rotated at angles of 10, 30, 60, 90, 180, and 270 degrees, images with inadequate lighting and noise, and images with optimal lighting and minimal noise. The findings of the experiment have been documented in Tables 5 and 6.

Table 5. Accumulative evaluation metric for PDF files

No	Test Scenario	Recall	Precision	F1
1	Normal orientation/position	0.974843	0.974843	0.974843
2	Rotated 10 degrees	0	0	0
3	Rotated 30 degrees	0	0	0
4	Rotated 60 degrees	0	0	0
5	Rotated 90 degrees	0.993827	0.98773	0.990769
6	Rotated 180 degrees	0	0	0
7	Rotated 270 degrees	0	0	0
8	Noisy data	0.13253	0.666667	0.221106
9	Less noise data	0.993827	0.98773	0.990769

Table 6. Accumulative evaluation metric for JPG files

No	Test Scenario	Recall	Precision	F1
1	Normal orientation/position	0.974843	0.974843	0.974843
2	Rotated 10 degrees	0	0	0
3	Rotated 30 degrees	0	0	0
4	Rotated 60 degrees	0	0	0
5	Rotated 90 degrees	0.888889	0.993103	0.938111
6	Rotated 180 degrees	0.888889	0.993103	0.938111
7	Rotated 270 degrees	0.888889	0.993103	0.938111
8	Noisy data	0.204819	1	0.34
9	Less noise data	0.889571	0.993151	0.938511

Tables 5 and 6 summarize the Recall, Precision, and F1 calculations from the results obtained from various previous testing scenarios. The values of these evaluation metrics indicate the performance of the OCR engine when applied to textbook document objects. The OCR system can effectively detect and recognize text in normal orientation or position, as this is consistent with findings from other studies [18, 19]. This occurred consistently for both PDF and JPG format data. High recall, precision, and F1 evaluation results indicate that OCR performance in these conditions is very good. These results are obtained for documents with normal or ideal image positions and good image quality [19]. This study found that image rotation at angles of 10, 30, and 60 degrees posed significant challenges for the OCR system, where no ISBN character values could be detected, as indicated by the low recall, precision, and F1 values. This challenge in handling documents with certain orientations has also been found and serves as a basis for another study [20]. Several studies have adopted various approaches to address this challenge, such as deep learning-based methods [21] and the adaptive circle projection algorithm [22]. Meanwhile, different results for upright rotations (90, 180, and 270 degrees) between PDF and JPG formats also produced unstable results. In PDF files, the OCR system could only detect text in images with a 90-degree rotation but failed to detect any characters with 180 and 270-degree rotations. This can be seen from the high F1 value for images with a 90-degree rotation. Meanwhile, in testing on JPG files, the OCR system successfully detected ISBN character values at all upright angles, namely 90, 180, and 270 degrees. The examination results using JPG data with 90, 180, and 270-degree rotations were the same, with high recall, precision, and F1 values.

Low lighting and noise negatively impact OCR performance [23], with low recall values, especially in PDF-formatted images. This indicates that further studies are needed to develop text recognition techniques that can overcome the challenges of low lighting and image noise. This means that in the area of text recognition, the quality of the image processing is essential. The results of tests on images with low lighting and noise showed quite low recall values in both PDF and JPG file tests. Although both test results showed higher precision values than recall, the low F1 value indicates that OCR has difficulty recognizing text in images with low lighting and much noise, whether in PDF or JPG files. In tests on PDF and JPG-formatted images with good lighting and less noise, high recall, precision, and F1 values were obtained. This indicates that OCR successfully recognized text well under these conditions. From the test results, OCR performance in recognizing ISBN text in textbook documents provides quite promising results, especially in ideal document conditions, both in image position and image and lighting quality. This provides an opportunity to apply OCR to automatically detect text content in documents, facilitating document-based evaluation processes. The findings of this research demonstrate that OCR technology can achieve strong performance under optimal conditions, as seen in scenarios with good lighting, minimal noise, and standard orientations, with high recall, precision, and F1 values. However, the results also highlight the system's limitations in handling challenging scenarios, such as images with low lighting, significant noise, or non-standard rotations (e.g., 10, 30, and 60 degrees), where performance significantly drops. These findings underscore the importance of improving OCR preprocessing techniques to enhance robustness across various conditions while maintaining its potential for practical applications, such as automated ISBN detection in textbook evaluation processes. Although the test results show that OCR successfully detects most ISBNs with high F1 values, some original characters remain unrecognized, and some characters are incorrectly detected. As seen in the tests of scanned images in both JPG and PDF formats, four original characters failed to be identified, and four characters were wrongly detected under normal conditions. In images rotated at 90, 180, and 270 degrees, 18 original characters failed to be identified, and one character was wrongly detected. In tests with poor lighting and noise, 11 characters were incorrectly detected. One character was wrongly detected in tests with good lighting and less noise. This indicates the need for further studies to explore improvement strategies to enhance character recognition accuracy in OCR.

4. CONCLUSION

In conclusion, the analysis of 90 book images under various conditions using the Tesseract OCR platform demonstrates Tesseract excels in text detection on high-quality, low-noise images, particularly those in normal orientation. The system performs notably well when processing JPG images with upright rotations of 90, 180, and 270 degrees. However, its effectiveness significantly diminishes when dealing with images rotated at 10, 30, and 60 degrees or with lower quality and higher noise. This variability in performance based on file type and image conditions highlights the importance of considering the specific characteristics of document images when using Tesseract OCR. This study has practical implications for applications such as mobile scanning and document digitization, where reliable text recognition is essential. However, the research boundaries are limited to specific conditions, including JPG and PDF formats, particular rotation angles, and controlled noise levels. Future research should focus on developing more advanced preprocessing methods and algorithms to improve Tesseract's capacity to accurately detect text in images that are rotated, poorly lit, or contain high levels of noise. Enhancing OCR performance under such challenging conditions will expand the applicability and reliability of Tesseract, thereby increasing its efficacy for practical applications and benefiting the public through improved accessibility to document digitization technology.

5. ACKNOWLEDGEMENTS

The authors would like to express their sincere gratitude to Politeknik Negeri Malang for the financial support. This support has been instrumental in facilitating the successful completion of this research. The authors also extend their appreciation to Library of Information Technology Department, Politeknik Negeri Malang, for providing access to textbook samples, which served as the primary data source for the experiment.

6. DECLARATIONS

AUTHOR CONTIBUTION

Conceptualization: Imam Fahrur Rozi, Endah Septa S, Astrifidha Rahma A. Methodology: Imam Fahrur Rozi, Arin Kistia N, Yuri Ariyanto Discussion of results: Imam Fahrur Rozi, Arin Kistia N. Writing Original Draft: Ahmadi Yuli A, Yuri Ariyanto, Ain Kistia N Writing Review and Editing: Imam Fahrur Rozi. Resources: Imam Fahrur Rozi, Ahmadi Yuli Ananta, Endah Septa S, Astrifidha Rahma A, Arin Kistia N Supervision: Imam Fahrur Rozi, Yuri Ariyanto. Approval of the final text: Imam Fahrur Rozi, Yuri Ariyanto.

FUNDING STATEMENT

This research was supported and funded by Politeknik Negeri Malang through the Penelitian Reguler Kompetisi scheme, organized by UPT P2M. The authors gratefully acknowledge this support, which has contributed to the successful completion of this study.

COMPETING INTEREST

The authors declare that there are no competing interests related to this research, its findings, or its publication.

REFERENCES

[1] S. M. Setiana, I. M. Yukasih, M. Dirgandini, and D. S. Halibanon, "The Effect of Lecturer Certification on Improving Teaching Performance: A Case Study of Japanese Language Lecturers in West Java," in *Proceeding of International Conference on Business, Economics, Social Sciences, and Humanities*. Universitas Komputer Indonesia, mar 2023, pp. 396–400, https://doi.org/10.34010/icobest.v4i.393.

[2] U. Rahardja, N. Lutfiani, A. Setiani Rafika, and E. Purnama Harahap, "Determinants of Lecturer Performance to Enhance Accreditation in Higher Education," in *2020 8th International Conference on Cyber and IT Service Management (CITSM)*. IEEE, oct 2020, pp. 1–7, https://doi.org/10.1109/CITSM50537.2020.9268871.

[3] A. F. Wulandari, A. Winarno, B. S. Luturlean, and F. Nur, "Explaining Gender in Moderating the Effect of Competency, Work Discipline and Job Satisfaction on Lecturer Performance," *Al-Tanzim: Jurnal Manajemen Pendidikan Islam*, vol. 8, no. 2, pp. 650–663, 2024, https://doi.org/10.33650/al-tanzim.v8i2.7193.

[4] F. Riandari, H. T. Sihotang, and H. Husain, "Forecasting the Number of Students in Multiple Linear Regressions," *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 21, no. 2, pp. 249–256, 2022, https://doi.org/10.30812/matrik.v21i2.1348.

[5] J. Memon, M. Sami, R. A. Khan, and M. Uddin, "Handwritten Optical Character Recognition (OCR): A Comprehensive Systematic Literature Review (SLR)," *IEEE Access*, vol. 8, pp. 142 642–142 668, 2020, https://doi.org/10.1109/ACCESS.2020.3012542.

[6] S. Drobac and K. Lindén, "Optical Character Recognition with Neural Networks and Post-Correction with Finite State Methods," *International Journal on Document Analysis and Recognition (IJDAR)*, vol. 23, no. 4, pp. 279–295, 2020, https://doi.org/10.1007/s10032-020-00359-9.

[7] R. M. Ahmed, T. A. Rashid, P. Fattah, A. Alsadoon, N. Bacanin, S. Mirjalili, S. Vimal, and A. Chhabra, "Kurdish Handwritten character recognition using deep learning techniques," *Gene Expression Patterns*, vol. 46, p. 119278, 2022, https://doi.org/10.1016/j.gep.2022.119278.

[8] M. Li, T. Lv, J. Chen, L. Cui, Y. Lu, D. Florencio, C. Zhang, Z. Li, and F. Wei, "TrOCR: Transformer-Based Optical Character Recognition with Pre-trained Models," in *The Thirty-Seventh AAAI Conference on Artificial Intelligence (AAAI-23)*, 2023, pp. 13 094–13 102, https://doi.org/10.48550/arXiv.2109.10282. [Online]. Available: www.aaai.org

[9] C. Clausner, A. Antonacopoulos, and S. Pletschacher, "Efficient and effective OCR engine training," *International Journal on Document Analysis and Recognition (IJDAR)*, vol. 23, no. 1, pp. 73–88, 2020, https://doi.org/10.1007/s10032-019-00347-8.

[10] S. Dome and A. P. Sathe, "Optical Charater Recognition using Tesseract and Classification," in *2021 International Conference on Emerging Smart Computing and Informatics (ESCI)*. IEEE, mar 2021, pp. 153–158, https://doi.org/10.1109/ESCI50559.2021.9397008.

[11] 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, https://doi.org/10.1007/s42001-021-00149-1.

[12] N. Anwar, T. Khan, and A. F. Mollah, "Text Detection from Scene and Born Images: How Good is Tesseract?" in *Recent Trends in Communication and Intelligent Systems*. Singapore: Springer, may 2022, pp. 115–122, https://doi.org/10.1007/978-981-19-1324-2_13.

[13] A. D. R N, S. Chinta, N. K. Ashili, B. S. Babu, R. R. Vydugula, and R. S. VSL, "An Intelligent Invoice Processing System Using Tesseract OCR," in *2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS)*. IEEE, apr 2024, pp. 1–6, https://doi.org/10.1109/ADICS58448.2024.10533509.

[14] A. Benaissa, A. Bahri, A. El Allaoui, and M. Abdelouahab Salahddine, "Build a Trained Data of Tesseract OCR engine for Tifinagh Script Recognition," *Data and Metadata*, vol. 2, p. 185, 2023, https://doi.org/10.56294/dm2023185.

[15] Tarun, T. Chauhan, and Varsha, "The Efficacy of Tesseract OCR: Insights from a Practical Application Study," in *11th International Conference on Cutting-Edge Developments in Engineering Technology and Science*. ICCDETS, may 2024, pp. 1601–1605, https://doi.org/10.62919/hdsg3874. [Online]. Available: https://cd69d634-fbf5-4870-9835-0d3668731238.filesusr.com/ugd/2b9d2a_fd5d4eaed0f643738391ad03ee00f821.pdf

[16] T. T. H. Nguyen, A. Jatowt, M. Coustaty, and A. Doucet, "Survey of Post-OCR Processing Approaches," *ACM Computing Surveys*, vol. 54, no. 6, pp. 1–37, jul 2022, https://doi.org/10.1145/3453476.

[17] D. Khairani, D. A. Bangkit, N. F. Rozi, S. U. Masruroh, S. Oktaviana, and T. Rosyadi, "Named-Entity Recognition and Optical Character Recognition for Detecting Halal Food Ingredients: Indonesian Case Study," in *2022 10th International Conference on Cyber and IT Service Management (CITSM)*. IEEE, sep 2022, pp. 01–05, https://doi.org/10.1109/CITSM56380.2022.9935966.

[18] L. Jianyang, B. Junrong, L. Bingjin, F. Zhiang, and Z. Su, "The Character Recognition Method Based on OCR," in *2023 26th ACIS International Winter Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD-Winter)*. IEEE, jul 2023, pp. 92–95, https://doi.org/10.1109/SNPD-Winter57765.2023.10223979.

[19] K. Olejniczak and M. Šulc, "Text Detection Forgot About Document OCR," in *CEUR Workshop Proceedings*. CEUR Workshop Proceedings, 2023. [Online]. Available: http://ceur-ws.org

[20] L. Jain, M. J. Wilber, and T. E. Boult, "Issues in Rotational (Non-)invariance and Image Preprocessing," in *2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, jun 2013, pp. 76–83, https://doi.org/10.1109/CVPRW.2013.19.

[21] P. Wang, J. Qiao, and N. Liu, "An Improved Convolutional Neural Network-Based Scene Image Recognition Method," *Computational Intelligence and Neuroscience*, vol. 2022, pp. 1–10, 2022, https://doi.org/10.1155/2022/3464984.

[22] J. He, Z. Zhang, H. Zhao, and J. Yang, "ACP- based Circular target image Rotation normalization system," in *2023 4th International Conference on Computer Vision, Image and Deep Learning (CVIDL)*. IEEE, may 2023, pp. 17–20, https://doi.org/10.1109/CVIDL58838.2023.10166580.

[23] D. Purwanto and A. Agustiyar, "Global Thresholding Implementation for Noise Handling in Digital Image Recognition," *Jurnal Transformatika*, vol. 21, no. 2, p. 93, 2024, https://doi.org/10.26623/transformatika.v21i2.8713.

[This page intentionally left blank.]