



Comparative Evaluation of Thresholding Methods for Optimized Digital Document Parsing Accuracy

Muhammad Noko Darpito¹, Kartika Firdausy², Abdul Fadlil³

¹Master Program of Informatics, Faculty of Industrial Technology, Ahmad Dahlan University, Yogyakarta, Indonesia

^{2,3}Electrical Engineering Study Program, Faculty of Industrial Technology, Ahmad Dahlan University, Yogyakarta, Indonesia

*2407048001@webmail.uad.ac.id

Abstract

Automated parsing of semi-structured documents has become increasingly important, particularly in standardized formats like SATS-LN, which contain fixed-layout fields such as permit number, addresses, validity period, and item types. This study investigates the impact of two thresholding methods Otsu and Sauvola on object detection accuracy using Faster R-CNN with Detectron2. A dataset of 200 SATS-LN documents, captured via scanner and camera, was augmented into 3,600 images and labeled for seven key fields. Image quality was evaluated using PSNR, SSIM, and MSE, while detection performance was measured through mAP, AP50, AP75, AR@100, precision, recall, and F1-score. Results showed that Sauvola preserved structural layout more effectively (SSIM: 0.76 for scanner, 0.47 for camera), although Otsu achieved higher PSNR on scanned images. Sauvola attained the highest macro and weighted F1-score (0.998), with near-perfect label detection and consistent performance across augmentations. Overall, Sauvola is more reliable for enhancing segmentation and detection in layout-based document processing.

Keywords: digitalization; document; image thresholding; parsing

How to Cite: Muhammad Noko Darpito, Kartika Firdausy, and Abdul Fadlil, "Comparative Evaluation of Thresholding Methods for Optimized Digital Document Parsing Accuracy", *J. RESTI (Rekayasa Sist. Teknol. Inf.)*, vol. 10, no. 2, pp. 453 - 464, Apr. 2026.

Permalink/DOI: <https://doi.org/10.29207/resti.v10i2.6982>

Received: July 20, 2025

Accepted: April 18, 2026

Available Online: April 26, 2026

This is an open-access article under the [CC BY 4.0 License](#)

Published by [Ikatan Ahli Informatika Indonesia](#)

1. Introduction

Currently, document digitization has become essential for effective and efficient information management. It improves accessibility, accelerates information retrieval and distribution, and reduces the risk of physical damage because digital data are safer from environmental hazards such as fire or flooding [1]. It also enables organizations to manage information better, improve services, and support deeper data tracking and analysis [1-2]. This process involves systematic stages, from data acquisition to document storage and management in accessible digital formats [1-2].

Otsu's method is a well-established global thresholding technique that determines an optimal threshold by minimizing intra-class variance and maximizing inter-class variance. It is particularly effective for images with clear foreground and background separation [3-5]. Previous studies have shown that Otsu can reduce noise

and improve image processing quality in various contexts, including medical imaging [6-7].

The Sauvola method is a local thresholding approach that is effective for handling brightness variation, particularly in degraded documents or images with poor lighting. Unlike Otsu, which applies a single threshold to the whole image, Sauvola calculates a threshold for each pixel based on the local mean and standard deviation of surrounding pixels. This enables more adaptive and accurate segmentation under uneven illumination [8-9].

Detectron2 is a PyTorch-based framework developed by Facebook AI Research for object detection and instance segmentation [10-11]. Its ability to detect and classify objects makes it relevant for document analysis and parsing, where document structures and contents often have complex layouts.

The SATS-LN document has a semi-structured format with a consistent layout, where important information is organized into fixed fields and table blocks. These

include permit number, transaction type, sender and recipient details, company addresses, validity period, ports of origin and destination, and specimen information such as scientific names, quantities, source, and quotas. Because of its structured format and fixed element positions, SATS-LN is suitable for document layout analysis and template-based object detection. However, scanned or photographed documents often contain uneven lighting, low resolution, and visual noise or distortion that can reduce extraction accuracy [12].

Despite the widespread use of thresholding in document image processing, most previous studies have focused on improving OCR performance or visual quality. However, its effect on object detection accuracy in structured document parsing, particularly for real-world layouts such as SATS-LN, remains underexplored. Official documents such as SATS-LN also present challenges due to inconsistent image quality, structured layouts, and multiple semantic zones. Therefore, this study evaluates the effectiveness of Otsu and Sauvola thresholding on the performance of a Faster R-CNN-based layout detection model using real SATS-LN documents.

Applying thresholding methods such as Otsu and Sauvola to document parsing enables clearer separation between text and background, which is important because binarization is a fundamental preprocessing step in document analysis systems [13]. Threshold-based techniques, including global and adaptive methods, also remain widely used in document binarization [14]. Therefore, this study is expected to show how thresholding affects digital document parsing before further processing stages.

2. Methods

This study was conducted to evaluate the effect of thresholding methods on the accuracy of data element detection in digital documents using a deep learning-based object segmentation approach. In this study, two popular thresholding methods, Otsu and Sauvola, were used to simplify the image background and clarify the information elements in the documents. Additionally, object detection model training was performed using the Detectron2 framework with the Faster R-CNN architecture and ResNet-50 backbone, running on the Google Colab platform with NVIDIA A100 GPU support to accelerate computation and model optimization.

This approach enables efficient evaluation of model performance under various visual input conditions generated by different thresholding methods. The entire methodology is illustrated in Figure 1.

2.1 Data Collection and Compilation

The initial stage of this study was the collection of a dataset of digital documents based on a template designed to resemble SATS-LN documents and using

dummy data representing the original elements of the documents.

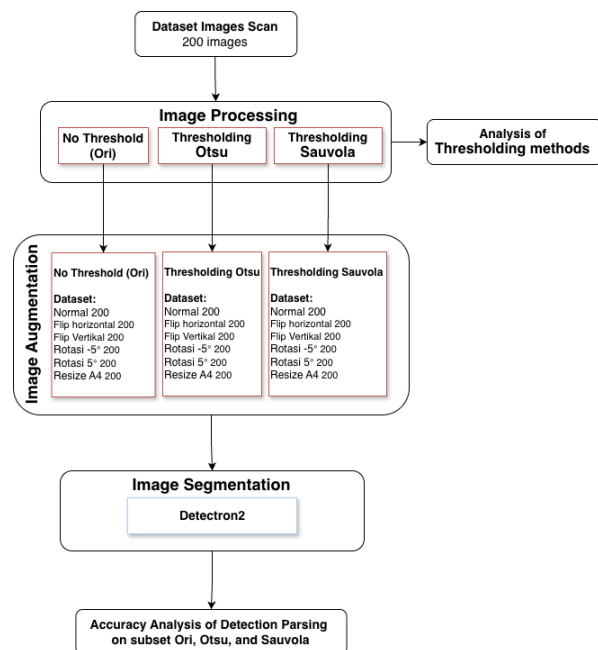


Figure 1. Research methodology

The dataset was printed and collected in the form of scanned document images, consisting of 200 PNG format document images. Of these, 50 images were taken using a camera smartphone (named cam1.png to cam50.png) as shown in Figure 2, while the remaining 150 images were obtained from scanning using a scanner (named doc1.png to doc150.png) as shown in Figure 3.

All documents are arranged in a uniform and consistent layout and contain important information that is the target of the detection process, such as document numbers, exporter and importer data, validity period, port of origin and destination, and specimen details.

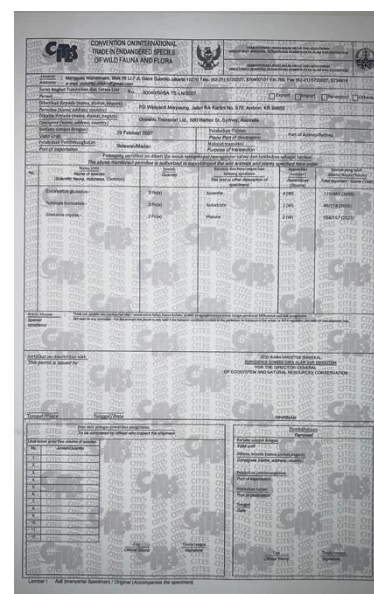


Figure 2. Example of a dataset scanned using a smartphone camera

The form is a permit for the export of specimens of endangered species. It includes the following sections:

- Header:** Convention on International Trade in Endangered Species of Wild Fauna and Flora.
- Permit Information:**
 - Permit No.: 00015/MNSA TS-LN/2025
 - Species: Dendrobium subtile
 - Quantity: 1 (one)
 - Origin: Perum Natsir Tsk, Rue Sainte-Catherine No. 375, Montreal, Canada
 - Destination: Port of Vancouver/Vancouver
- Export/Import Information:**
 - Export/Import: ☒ Export ☐ Import ☐ Re-export ☐ Others
 - Permit No.: 00015/MNSA TS-LN/2025
- Specimen Information:**
 - Species: Dendrobium subtile
 - Quantity: 1 (one)
 - Origin: Perum Natsir Tsk, Rue Sainte-Catherine No. 375, Montreal, Canada
 - Destination: Port of Vancouver/Vancouver
- Signature and Date:**
 - Signature: [Signature]
 - Date: 13 Desember 2022

Figure 3. Example of a dataset scanned using a scanner

2.2 Image Thresholding Process

After dataset collection, image segmentation was performed using Otsu and Sauvola thresholding, producing three subsets: ori, Otsu, and Sauvola from 200 document images.

This stage aimed to simplify the background and improve document clarity for the elements to be annotated and detected. By reducing noise and uneven illumination, preprocessing enabled clearer foreground-background separation and supported more accurate feature extraction for object detection. Thresholding was applied before training, with each original image processed under three conditions: without thresholding (ori), global Otsu, and adaptive Sauvola. These outputs were treated as separate input datasets for Faster R-CNN training in Detectron2.

The pipeline consisted of grayscale conversion, thresholding, data augmentation, annotation transformation, and dataset splitting into training, validation, and testing sets. Each subset was then used to train a separate Faster R-CNN model with identical architecture and hyperparameters, ensuring that performance differences reflected thresholding effects rather than model configuration.

2.3 Thresholding Result Analysis

The thresholding results are then evaluated to measure the quality of the processing using three quantitative

metrics, namely Mean Squared Error (MSE), Peak Signal-to-Noise Ratio (PSNR), and Structural Similarity Index Measure (SSIM). This analysis aims to compare the clarity and structural similarity of the thresholding results with the original image, so that the thresholding method that provides the best visual quality can be determined as a basis for the subsequent segmentation process. MSE is a metric used to measure the average square error between the pixels of the processed image and the reference image. The lower the MSE value, the smaller the difference between the two images. The MSE formula is defined as Equation 1.

$$MSE = \frac{1}{MN} \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} (I(i,j) - K(i,j))^2 \quad (1)$$

$I(i,j)$ is the pixel intensity at position (i,j) of the original image, $K(i,j)$ is the pixel intensity in the thresholding image, and M and N are the image dimensions.

MSE measurement is not limited to evaluating image segmentation results but is also widely applied in various other fields such as digital watermarking, where the performance of various algorithms is evaluated using MSE along with other metrics such as PSNR and SSIM [15]. This broader application demonstrates the flexibility of MSE as a standard measure for assessing the fidelity of an image to its original. PSNR is a metric used to measure the clarity of the processed image relative to the original image. PSNR is the ratio between the maximum pixel value and the noise level. A higher PSNR value indicates better visual quality [16]. PSNR is calculated using Equation 2.

$$PSNR = 10 \cdot \log_{10} \left(\frac{MAX^2}{MSE} \right) \quad (2)$$

MAX is the maximum pixel value of the image (usually 255 for 8-bit images), and MSE is the result of the previous formula.

SSIM is a metric that evaluates the similarity of local structures between two images. SSIM considers three aspects: luminance, contrast, and structure. SSIM values range from -1 to 1, with a value of 1 indicating that the two images are structurally identical [17-18]. The SSIM formula is generally expressed as Equation 3.

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \quad (3)$$

μ_x and μ_y are the average pixel intensities of images x and y , σ_x^2 and σ_y^2 are the variances of each image, and σ_{xy} is the covariance between the two images. The constants C_1 and C_2 are used to avoid division by zero.

By using these three metrics together, the quality of the thresholding results can be evaluated more comprehensively in terms of pixel error (MSE), visual clarity (PSNR), and structural similarity (SSIM).

2.4 Document Labeling

After thresholding and image quality analysis, manual annotation was performed on 200 document images in

their normal position (without augmentation). Seven main document elements were labeled using rectangular bounding boxes, namely *number*, *exporter_name_address*, *destination_name_address*, *expiry_date*, *port_of_origin*, *port_of_destination*, and *type_of_goods*. These annotations were then used as references to automatically generate labels for augmented images by adjusting bounding box coordinates to geometric transformations such as rotation, flipping, and scaling. This process ensured that the annotations remained accurate and consistent with the positions of document elements after visual transformation.

Image processing techniques emphasize the importance of accuracy when handling document images to facilitate reliable machine learning applications in document analysis [19]. Therefore, ensuring that coordinate adjustments are performed carefully is essential to improve the system's robustness in identifying and extracting relevant information from enhanced document images.

2.5 Image Augmentation Process

To enrich the variety of training data and improve the model's resilience to changes in orientation and scale, an augmentation process was performed on each image subset, namely *ori*, *otsu*, and *sauvola*. The types of augmentation applied include rotation by $+5^\circ$, rotation by -5° , horizontal flipping (horizontal flip), vertical flipping (vertical flip), and resizing to A4 paper format. Each type of augmentation is applied to 200 images, increasing the total number of images in each subset to 1,200 images. Thus, this augmentation process produced a total of 3,600 images used in the entire experiment. A total of 3,600 images are generated from three subsets, namely *ori*, *otsu*, and *sauvola*, each comprising 1,200 images.

The application of augmentation techniques in Detectron2 has been shown to improve object detection performance. Its flexible framework allows various image transformations to be integrated easily into the training pipeline. Previous studies reported that augmentation can improve precision, recall, and F1-score in object detection tasks [20]. In addition, object-focused augmentation that incorporates contextual information has been shown to improve accuracy in complex visual environments [21]. These findings indicate that augmentation not only increases the amount of training data but also improves data representation quality.

In this study, augmentation was not intended to replace the original dataset, but to extend the representation of document images through controlled geometric transformations. This is important because the dataset consists of 200 original SATS-LN document images with a uniform layout and seven predefined data elements as detection targets. Therefore, augmentation was used to support a controlled comparative evaluation of thresholding methods under a consistent document

structure, rather than to establish broad generalization across highly diverse document types.

The selected augmentation techniques, namely rotation, flipping, and resizing, were chosen to simulate moderate variations during document acquisition, such as changes in orientation, page position, and scale. However, this process mainly represents controlled geometric variability and does not fully capture real-world conditions such as severe blur, strong shadows, folds, occlusion, or complex background noise. Thus, the augmented dataset should be interpreted as an extended representation of structured document variation rather than a comprehensive representation of all real-world document scenarios.

The annotation positions of document data elements were also adjusted automatically for each transformed image by recalculating bounding box coordinates according to the applied geometric transformation. In rotation, the coordinates were updated based on the rotation angle, while in flip and resize transformations, the annotation position and scale were adjusted to remain consistent with the actual element location. This process is important to ensure that all training data, including images and annotations, remain representative and valid for the visual conditions encountered during model training.

2.6 Training of Detection Models with Detectron2

The augmented images and adjusted annotations were used as training data to build object detection models. Training was conducted using the Faster R-CNN architecture with a ResNet-50 C4 backbone implemented in the Detectron2 framework. The model was trained separately for each subset, namely *ori*, *otsu*, and *sauvola*.

In this study, the 200 original document images were first divided into training, validation, and testing sets with a ratio of 80%, 10%, and 10%, resulting in 160 original documents for training, 20 for validation, and 20 for testing. Augmentation was then applied independently within each partition, producing 960 training images, 120 validation images, and 120 testing images in each subset. The same split configuration was consistently applied to the *ori*, *otsu*, and *sauvola* subsets to maintain comparability between thresholding methods and ensure evaluation under the same experimental setting.

This training design was intended to ensure that comparison between subsets reflects the effect of image preprocessing under consistent model settings. The objective was to enable the model to accurately and consistently recognize and detect the seven document data elements across various image variations.

2.7 Detection Accuracy Analysis with Detectron2

The final step of this study was to evaluate object detection models trained on the *ori*, *Otsu*, and *Sauvola* subsets. Evaluation used mean average precision (mAP)

[22], together with AP50 and AP75, to measure detection accuracy at different intersection over union (IoU) thresholds. APs, APm, and API were also used to assess performance on small, medium, and large objects, providing a detailed view of model robustness to scale variation [23]. To strengthen the evaluation, Average Recall at 100 proposals (AR@100), accuracy, and macro precision were included to measure recall capacity, overall correctness, and balanced class-wise precision across all labels [24-25].

In addition, Macro Recall, Macro F1 Score, Weighted Precision, Weighted Recall, Weighted F1 Score, and Support were used to provide a more comprehensive view of the model's classification ability. Macro Recall and Macro F1 Score assess performance evenly across all classes, including minority classes [26-27], whereas weighted metrics consider class proportions and are useful when class distribution is uneven [28-29]. Support indicates the number of samples for each label and helps interpret metric reliability; for example, very low support may increase the risk of inaccurate detection in a class such as item type [30].

Precision, recall, and F1-score were also measured to evaluate the balance between true positive and false

detections. All evaluations were conducted separately for each subset and across all seven document element classes. Together, these metrics form the basis for assessing the effectiveness of each thresholding method in supporting automatic segmentation and information detection in digital documents.

3. Results and Discussions

3.1 Image Thresholding Process and Result

Before conducting object detection model training, document images were first processed using two thresholding methods namely Otsu and Sauvola, and a subset of images without thresholding (ori) was also included as a comparison.

The dataset used consists of a combination of images captured using a camera (cam) as shown in Figure 4 and scanned using a scanner (doc) as shown in Figure 5. Camera images tend to have uneven lighting, low contrast, and perspective irregularities, while scanner images are generally cleaner, more focused, and have more stable visual quality.

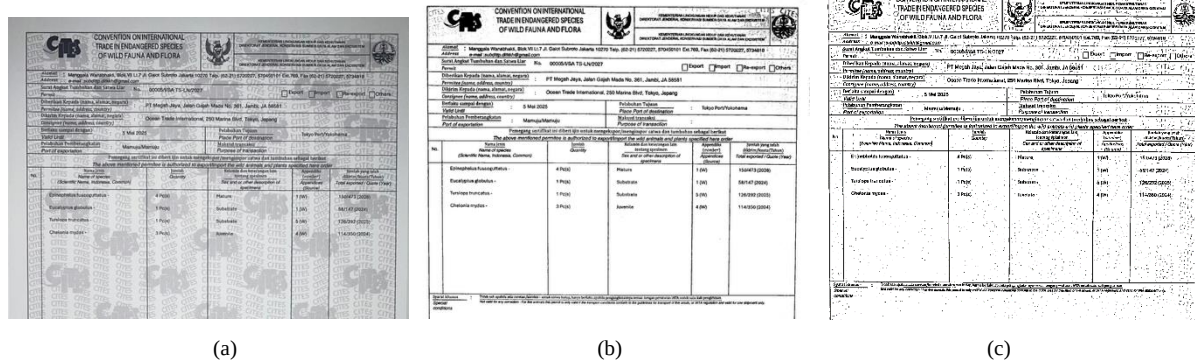


Figure 4. Thresholded image on scanned documents using a camera: (a) No threshold, (b) Otsu, (c) Sauvola



Figure 5. Thresholded image on scanned documents using a scanner: (a) No threshold, (b) Otsu, (c) Sauvola

To determine the visual quality of the images produced by each method, an evaluation was conducted using three quantitative metrics, namely Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), and Mean Squared Error (MSE).

Table 1 shows a comparison of the average PSNR, SSIM, and MSE values of each thresholding method for

50 camera-scanned document images and 150 scanner-scanned document images. The PSNR value is used to measure the clarity of the thresholded image relative to the original image, SSIM measures the similarity of local structures between the two images, and MSE measures the average error between pixels. These values provide a comprehensive overview of how well the thresholding method preserves important

information in the document while filtering out unnecessary background.

Table 1. Evaluation Thresholding on Camera and Scanner Images

File type	Otsu			Sauvola		
	PSNR	SSIM	MSE	PSNR	SSIM	MSE
cam	11.05	0.45	5122.42	10.91	0.47	5281.55
doc	23.58	0.73	285.28	19.95	0.76	661.65

Evaluation of camera-acquired images shows that thresholding quality was strongly affected by the method used. Otsu produced a PSNR of 11.05, while Sauvola produced 10.91, indicating that both methods faced similar difficulty in maintaining visual clarity. In terms of SSIM, Sauvola performed slightly better at 0.47 than Otsu at 0.45, suggesting better preservation of local structures such as text and box lines. However, Otsu achieved a lower MSE of 5122.42 compared with 5281.55 for Sauvola, indicating better global pixel stability. Thus, Sauvola was better in preserving local structure, whereas Otsu was more stable at the pixel level for camera-based images.

The relatively low SSIM values in camera-based documents indicate substantial structural deviation from the original images. Compared with scanner-acquired images, camera images are more affected by non-uniform illumination, shadow gradients, perspective distortion, and sensor noise, which produce uneven intensity distributions. Although Sauvola preserved local structures better than Otsu under varying brightness, both methods remained sensitive to extreme lighting imbalance.

This limitation reflects a basic challenge in real-world document parsing. Thresholding mainly performs intensity-based segmentation and does not explicitly address illumination normalization or shading correction. As a result, structural similarity may still decrease even when foreground text remains visible. These findings indicate that additional preprocessing, such as illumination correction or contrast normalization, may be needed to improve robustness for camera-acquired documents.

In scanner-acquired images, the performance difference between Otsu and Sauvola was more evident. Otsu achieved a PSNR of 23.58, the highest among all image and method combinations in this study, indicating high visual clarity with minimal noise. In contrast, Sauvola obtained a PSNR of 19.95, showing that its local approach was less effective than Otsu in preserving clarity on relatively clean scanned images.

The SSIM values for scanner images were also close, with Otsu at 0.73 and Sauvola at 0.76. This indicates that both methods were able to preserve the overall document structure, although Sauvola showed slightly better local structural similarity.

These results suggest that Otsu is more suitable for documents with relatively uniform pixel intensity, as commonly found in scanned images. In terms of MSE,

Otsu again performed better, with a value of 285.28 compared with 661.65 for Sauvola.

This result confirms that Otsu not only preserved image clarity and structure, but also minimized global pixel changes. Therefore, Otsu can be considered more efficient and stable for scanned images. Its superiority across scanned-image metrics strengthens the view that global thresholding is more suitable for high-quality images with minimal visual disturbance.

Although the comparison between Otsu and Sauvola showed observable differences in PSNR, SSIM, and MSE for both camera- and scanner-acquired images, the analysis remains descriptive. The values in Table 1 represent average results for each domain and are intended to show empirical tendencies in image quality. Because formal statistical significance testing has not yet been conducted, these findings should be interpreted as trends within the evaluated dataset rather than conclusive evidence of the absolute superiority of one method.

Future studies may strengthen this comparison by conducting statistical tests on PSNR, SSIM, and MSE at the individual-image level, since both thresholding methods were applied to the same document images. Such analysis would provide stronger evidence on whether the observed differences between Otsu and Sauvola remain consistent beyond the current experimental setting.

The relationship between image quality metrics and detection performance must therefore be interpreted carefully. In this study, PSNR and MSE mainly represent pixel-level fidelity between thresholded and original images, whereas SSIM reflects preservation of local structural information. Therefore, higher PSNR or lower MSE does not necessarily lead to better detection performance. This is reflected in the results, where Otsu achieved higher PSNR and lower MSE on scanned documents, Sauvola showed slightly better SSIM, and the ori subset still achieved the highest mAP in object detection. These findings indicate that different metrics represent different aspects of image representation: PSNR and MSE describe global similarity, whereas SSIM is more closely related to structural readability that supports localization and classification.

3.2 Object Detection Performance Evaluation Results

This study used the Detectron2 framework with Faster R-CNN and a ResNet-50 backbone to evaluate how different thresholding-based image representations affect the detection of document data elements. Thresholding was applied before model training so that each model learned from one specific input condition, namely ori, otsu, or sauvola. Therefore, the comparison in this section reflects the effect of image preprocessing on detection performance rather than differences in model architecture.

The three subsets were derived from the same original document collection and evaluated using the same

training configuration. After augmentation, each subset contained 1,200 images and was divided into training, validation, and testing sets using the same ratio across all experimental conditions. The training configuration used a batch size of 24 images per iteration (IMS_PER_BATCH = 24), 2,000 total iterations (MAX_ITER = 2000), and an initial learning rate of 0.001 (BASE_LR = 0.001). These hyperparameters were selected to balance convergence stability and computational efficiency in Google Colab with NVIDIA A100 GPU support. Model evaluation was based on mean Average Precision (mAP), including mAP, AP50, AP75, AR@100, accuracy, and macro precision, with the results presented in Table 2.

Table 2. Results of the Object Detection Model Performance

Subset	mAP	AP50	AP75	AR100	Acc	MPrec
Ori	0.79	0.94	0.81	0.90	0.99	0.99
Otsu	0.74	0.99	0.91	0.88	0.99	0.98
Sauvola	0.74	0.99	0.91	0.89	0.99	0.99

Based on the model trained on the ori subset, the best overall performance was achieved with a mAP of 0.79, indicating high precision across Intersection over Union (IoU) thresholds from 0.5 to 0.95. The AP50 and AP75 values of 0.94 and 0.81 show that the model detected objects accurately under both low and high overlap tolerances. The model also recorded an AR@100 of 0.90, indicating high sensitivity in capturing objects comprehensively. In classification, both accuracy and macro precision reached 0.99, showing balanced predictions across all label classes without domination by majority classes.

Meanwhile, the model trained on the Otsu subset produced a mAP of 0.74, with AP50 and AP75 values of 0.99 and 0.91. These values exceeded those of the ori subset in terms of detection precision at higher thresholds, indicating that Otsu effectively sharpened document object boundaries and supported highly precise predictions. However, the AR@100 value of 0.88 suggests a slight decrease in overall detection range, possibly due to the loss of contextual information caused by binary segmentation. Even so, the accuracy and macro precision remained high at 0.99 and 0.98, indicating stable classification despite minor variation across classes.

In the Sauvola subset, performance was almost identical to Otsu, with a mAP of 0.74, AP50 of 0.99, and AP75 of 0.91. These results indicate that adaptive thresholding was also effective in strengthening local features and improving predictions for objects with complex contours. Its AR@100 value of 0.89 was slightly higher than Otsu, suggesting that Sauvola preserved local structure better and made the model more sensitive to small or fragmented objects. The accuracy and macro precision values of 0.99 further show that this subset maintained both detection sensitivity and consistent classification across labels.

Overall, although the highest mAP was achieved by the ori subset, the Otsu and Sauvola subsets showed

superior high-threshold precision through AP50 and AP75, which is important in applications requiring high accuracy. In addition, Sauvola demonstrated a more balanced performance between precision and detection range.

Additionally, object detection performance was further analyzed using more specific classification metrics, namely macro recall, macro F1-score, weighted precision, weighted recall, and weighted F1-score. This evaluation provides a deeper view of the model's consistency and balance in recognizing all labels. These metrics are especially relevant for detecting semi-structured digital document elements with proportional label distribution. The classification results based on these metrics are presented in Table 3.

Table 3. Results of Model Classification Evaluation

Subset	Macro		Weighted		Support
	rec	F1	prec	rec	F1
Ori	0.997	0.996	0.995	0.996	0.996
Otsu	0.992	0.988	0.992	0.995	0.988
Sauvola	0.998	0.998	0.998	0.998	0.998

In the ori subset, the model showed very high and stable performance across all metrics, with a macro recall of 0.997 and a macro F1-score of 0.996. The weighted precision, weighted recall, and weighted F1-score were 0.995, 0.996, and 0.996, indicating that performance remained consistent even after considering class distribution.

These results support the previous mAP and AP findings that the ori subset provided balanced and high-quality detection and classification. The otsu subset also performed well, although slightly lower, with a macro recall of 0.992 and a macro F1-score of 0.988. This indicates that only a small proportion of objects may not have been classified correctly, while overall balance between precision and recall was still maintained. Its weighted precision, weighted recall, and weighted F1-score were 0.992, 0.995, and 0.988, respectively.

The slight decrease in the F1-score suggests that Otsu thresholding may reduce visual context in some document areas that are important for classification. In contrast, the Sauvola subset produced the highest results, with macro recall and macro F1-score both reaching 0.998, and weighted precision, weighted recall, and weighted F1-score also at 0.998.

This performance indicates that Sauvola thresholding was effective in preserving visual structure, allowing the model to recognize document elements with very high accuracy and consistency across classes. These values support the conclusion that, although thresholding changes image appearance, Sauvola can strengthen visual boundaries without reducing classification balance.

Although the Sauvola subset achieved the highest macro and weighted F1-scores of 0.998, this result should be interpreted in the context of the experimental setting. The dataset consisted of fixed-layout SATS-LN

documents with seven predefined element classes, making the task more constrained than general document layout analysis. Moreover, the ori subset also achieved a very high macro F1-score of 0.996. Therefore, the superiority of Sauvola reflects stronger classification stability under the current dataset conditions rather than direct evidence of broader generalization to more heterogeneous layouts.

The identical support value of 840 across all subsets indicates a fair comparison, with differences in macro recall, F1-score, and weighted metrics mainly reflecting preprocessing effects rather than sample size variation.

Overall, these evaluation results reinforce the previous findings that the ori subset excelled in overall detection performance, while Sauvola showed the best stability and classification balance.

Although the ori subset achieved the highest mAP of 0.79 and Sauvola produced the highest macro and weighted F1-scores of 0.998, these findings are not contradictory. Faster R-CNN with a ResNet-50 backbone relies on gradient transitions and texture information for accurate region proposal and bounding box regression. Because the ori subset preserves full grayscale intensity, it provides richer spatial features that improve localization across varying IoU thresholds and lead to higher mAP.

In contrast, thresholding simplifies grayscale gradients into binary foreground-background regions. While this improves structural clarity, it also removes fine texture details that support precise localization, which explains the slight decrease in mAP. However, Sauvola's adaptive local thresholding better preserves structural boundaries and layout consistency, leading to more stable and balanced per-label classification. These findings therefore indicate a trade-off between texture-preserving localization accuracy in ori and structure-enhanced classification stability in Sauvola, depending on the system objective.

The present experimental design can be regarded as a controlled comparison of thresholding conditions because the same document structure, annotation scheme, model architecture, and training configuration were maintained across the ori, otsu, and sauvola subsets. Under this setting, the main difference among the three experiments was the image representation generated by preprocessing. Therefore, the observed differences in detection and classification performance can reasonably be interpreted as reflecting the contribution of thresholding under a consistent experimental framework.

Nevertheless, this study does not yet include a broader staged ablation design to isolate all contributing factors more specifically, such as experiments without augmentation, experiments separated by acquisition domain during training, or additional preprocessing combinations. For this reason, the findings should be interpreted as evidence from a controlled comparative

setup rather than from a full ablation study. Further research may strengthen this aspect by including more detailed ablation experiments to quantify the individual contribution of each preprocessing component to detection performance.

A portion of this study focused on an in-depth evaluation of object detection performance using a label-based approach to measure how accurately the model identified each document element. This evaluation is important because each label has distinct visual and spatial characteristics that may affect detection results. Therefore, true positive (TP), false positive (FP), false negative (FN), precision, recall, and F1-score were measured separately for each label in each preprocessing subset. The results are presented in Table 4.

Based on Table 4, the ori subset shows excellent and stable detection performance. All labels except two achieved 100% precision, recall, and F1-score, indicating no classification errors. A slight decrease occurred in *inidestination_name_address* and *port_of_destination*. For *destination_name_address*, precision was 99%, recall 98%, and F1-score 98%, caused by 1 false positive and 3 false negatives. For *port_of_destination*, precision was 98% and recall 100%, resulting in an F1-score of 99% due to 3 false positives. Overall, this subset demonstrates very high detection precision and sensitivity, with only minor errors in two classes.

The Otsu subset also showed competitive results. Most labels were detected with high accuracy, although slight decreases appeared in several classes. *The number* and *exporter_name_address* labels each recorded 98% precision, 100% recall, and 99% F1-score, caused by 2 and 3 false positives. The *destination_name_address* label recorded 1 false positive and 2 false negatives, with 99% precision, 98% recall, and 99% F1-score. The *type_of_goods* label showed a slightly lower recall of 98% and an F1-score of 99% due to 2 false negatives, although precision remained 100%. Other labels, such as *expiry_date*, *port_of_origin*, and *port_of_destination*, continued to show maximum or near-perfect performance. These results indicate that Otsu thresholding can sharpen visual element boundaries, but some errors still occur in text structure or small elements.

The Sauvola subset recorded the most stable and overall best performance. Of the seven evaluated labels, six achieved 100% precision, recall, and F1-score without false positives or false negatives. The only exception was the number label, which recorded 98% precision, 100% recall, and 99% F1-score due to 2 false positives. This means that, despite minor additional prediction errors, all objects in that class were still detected correctly. The stability of the *destination_name_address* label, which was previously more error-prone in the ori and Otsu subsets, indicates that Sauvola is highly effective in preserving the

structural details of document elements through adaptive thresholding.

Table 4. Results of Object Detection Evaluation per Label

Subset	Label	TP	FP	FN	Prec (%)	Rec (%)	F1 (%)
Ori	number	120	0	0	100	100	100
	exporter_name_address	120	0	0	100	100	100
	destination_name_address	120	1	3	99	98	98
	expiry_date	120	0	0	100	100	100
	port_of_origin	120	0	0	100	100	100
	port_of_destination	117	3	0	98	100	99
	type_of_goods	120	0	0	100	100	100
Otsu	number	118	2	0	98	100	99
	exporter_name_address	120	3	0	98	100	99
	destination_name_address	121	1	2	99	98	99
	expiry_date	120	0	0	100	100	100
	port_of_origin	120	0	0	100	100	100
	port_of_destination	117	3	0	98	100	99
	type_of_goods	118	0	2	100	98	99
Sauvola	number	118	2	0	98	100	99
	exporter_name_address	120	0	0	100	100	100
	destination_name_address	120	0	0	100	100	100
	expiry_date	120	0	0	100	100	100
	port_of_origin	120	0	0	100	100	100
	port_of_destination	120	0	0	100	100	100
	type_of_goods	120	0	0	100	100	100

3.3 Image Thresholding Process and Result

Regarding the bounding box prediction results. This visualization was performed on augmented images from the Ori, Otsu and Sauvola thresholding subsets, which include transformation conditions such as horizontal flip, vertical flip, resize to A4, and normal position (without transformation). The purpose of this approach is to demonstrate how the model recognizes and localizes important elements of documents under various visual conditions.

As shown in the prediction visualizations for the Otsu subset in Figure 6, Sauvola in Figure 7, and Ori in Figure 8, the model successfully detected the seven target document elements such as *number*, *exporter_name_address*, *exporter_name_address*,

expiry_date, *port_of_origin*, *port_of_destination*, and *type_of_goods*, both on horizontally and vertically flipped images. The blue bounding box represents the ground truth (GT), while the red represents the model prediction (PD). The positional consistency between the two indicates that the model is able to maintain spatial accuracy even when the document undergoes geometric transformations.

Overall, these results indicate that the trained object detection model can reliably handle variations in document appearance and orientation. This visual evaluation supports the numerical results and demonstrates the system's potential for real-world document applications.

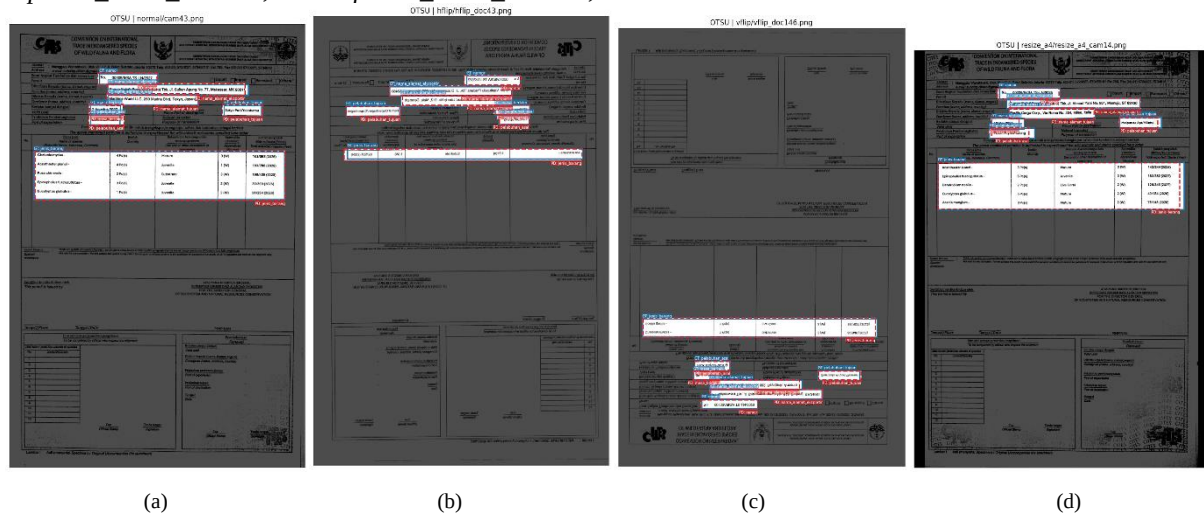


Figure 6. Visualization of object detection model predictions on the Otsu subset:
(a) Normal, (b) Horizontal flip, (c) Vertical flip, (d) Resize to A4

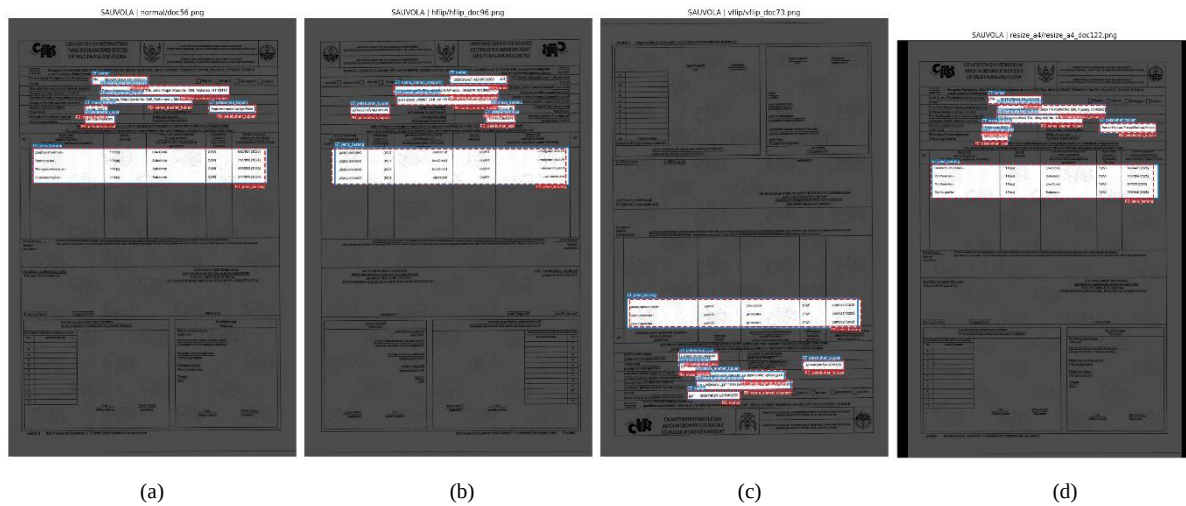


Figure 7. Visualization of object detection model predictions on the Sauvola subset:
(a) Normal, (b) Horizontal flip, (c) Vertical flip, (d) Resize to A4

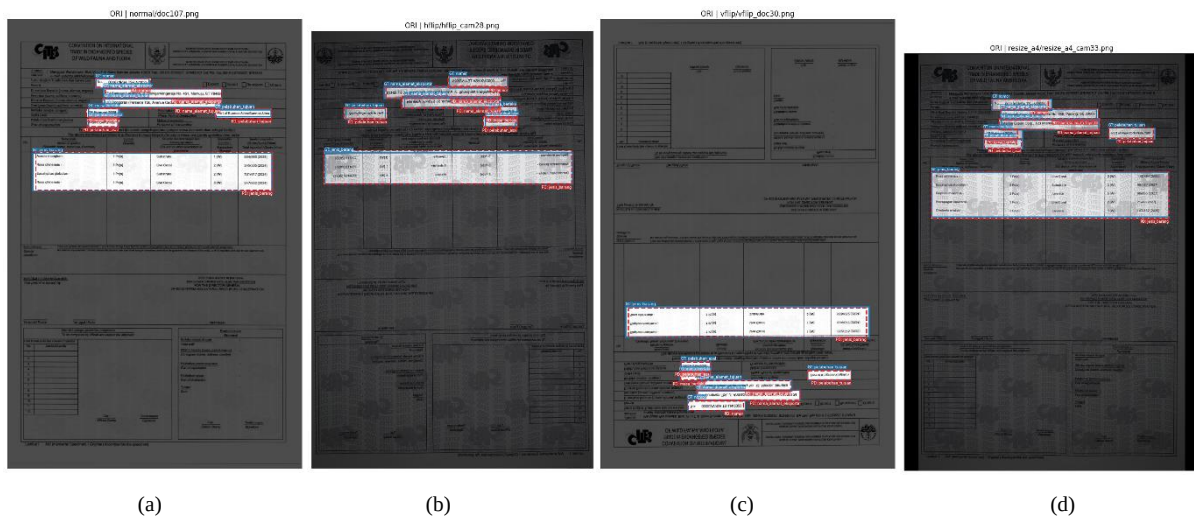


Figure 8. Visualization of object detection model predictions on the Ori subset:
(a) Normal, (b) Horizontal flip, (c) Vertical flip, (d) Resize to A4

Overall, these results indicate that the trained object detection model can reliably handle variations in document appearance and orientation. This visual evaluation supports the numerical results and demonstrates the system's potential for real-world document applications.

4. Conclusions

Based on the evaluation of thresholding effects on document data element detection using Faster R-CNN in Detectron2, three subsets were tested: ori, Otsu, and Sauvola. The evaluation covered image quality, object detection accuracy, label-wise classification, and visual analysis of bounding box predictions.

In image quality evaluation, Sauvola preserved document structure better than Otsu. On scanned documents, Sauvola achieved SSIM 0.76, PSNR 19.95, and MSE 661.65, while Otsu achieved SSIM 0.73, PSNR 23.58, and MSE 285.28. Although Otsu produced higher PSNR and lower MSE, Sauvola's higher SSIM indicates better structural preservation. On

camera-based documents, Sauvola also showed slightly better structural quality with SSIM 0.47, compared with 0.45 for Otsu. Overall, Sauvola generated more structurally reliable segmented images for downstream detection.

For object detection accuracy, the ori subset achieved the highest mAP of 0.79, while Otsu and Sauvola both recorded 0.74. The ori subset also reached AP50 0.94, AP75 0.81, and AR@100 0.90, indicating the strongest overall detection stability.

In contrast, Sauvola performed best in inter-label classification, with macro F1-score, weighted F1-score, and accuracy all reaching 0.998. Otsu also performed well, with macro recall 0.992 and macro F1-score 0.988.

Label-wise evaluation showed that the ori subset detected most document elements perfectly, with only minor errors in `destination_name_address` and `port_of_destination`, whose F1-scores slightly decreased to 98% and 99%. Sauvola produced the most

stable label-wise performance, with six of seven labels reaching perfect detection, while the number label showed a small decrease to 98% precision and 99% F1 due to two false positives. Thus, Sauvola maintained the highest classification consistency across labels.

Visual analysis of bounding box predictions also showed consistent performance under augmentation conditions such as normal, horizontal flip, vertical flip, and resize A4. Sauvola produced highly precise bounding boxes aligned with ground truth, whereas ori remained visually stable in the normal condition and Otsu showed slight shifts under more extreme transformations. Overall, ori gave the best overall detection, while Sauvola gave the most stable inter-label classification.

These findings are partly consistent with previous studies. Earlier research reported that Sauvola provided better visual clarity than Otsu in noisy documents, which agrees with the SSIM-based structural preservation found in this study. However, unlike prior studies that mainly focused on OCR or text segmentation, this study specifically examines the effect of thresholding on object detection accuracy through layout-aware bounding box analysis. This extends thresholding evaluation beyond OCR-centered metrics.

Overall, the results indicate that preprocessing should be selected according to system objectives. Under the present setting, ori produced stronger overall detection performance, whereas Sauvola produced more consistent classification. Future studies may explore adaptive or hybrid preprocessing strategies that combine ori-like processing and Sauvola thresholding based on input characteristics, while integrating OCR for end-to-end document extraction.

References

- [1] M. D. S. Saraswati and N. N. S. A. Yanti, "Manajemen Dokumen Digital Dalam Mendukung Layanan Informasi Di PT Air Minum Giri Menang (Perseroda)," *Guna Sewaka*, vol. 5, no. 1, pp. 16–26, Feb. 2025, doi: [10.53977/jgs.v5i1.2395](https://doi.org/10.53977/jgs.v5i1.2395).
- [2] I. G. Suputra Widharma, I. G. N. Sangka, I. N. Sunaya, I. M. Sajayasa, and I. G. K. Sri Budarsa, "Proses Digitalisasi Dokumen Laboratorium Dengan Menggunakan Omnipage Software," *Jurnal Ilmiah Vastuwidya*, vol. 6, no. 1, pp. 78–83, Feb. 2023, doi: [10.47532/jiv.v6i1.796](https://doi.org/10.47532/jiv.v6i1.796).
- [3] Z. Jing and B. Tang, "Improved image segmentation method based on Otsu thresholding and level set techniques," *J. Phys. Conf. Ser.*, vol. 2813, no. 1, p. 012017, Aug. 2024, doi: [10.1088/1742-6596/2813/1/012017](https://doi.org/10.1088/1742-6596/2813/1/012017).
- [4] S. Singh, N. Mittal, H. Singh, and D. Oliva, "Improving the segmentation of digital images by using a modified Otsu's between-class variance," *Multimed. Tools Appl.*, vol. 82, no. 26, pp. 40701–40743, Nov. 2023, doi: [10.1007/s11042-023-15129-y](https://doi.org/10.1007/s11042-023-15129-y).
- [5] J. Xing, P. Yang, and L. Qingge, "Robust 2D Otsu's Algorithm for Uneven Illumination Image Segmentation," *Comput. Intell. Neurosci.*, vol. 2020, pp. 1–14, Aug. 2020, doi: [10.1155/2020/5047976](https://doi.org/10.1155/2020/5047976).
- [6] T. Eelbode et al., "Optimization for Medical Image Segmentation: Theory and Practice When Evaluating With Dice Score or Jaccard Index," *IEEE Trans. Med. Imaging*, vol. 39, no. 11, pp. 3679–3690, Nov. 2020, doi: [10.1109/TMI.2020.3002417](https://doi.org/10.1109/TMI.2020.3002417).
- [7] K. Wang, Y. Pu, Y. Zhang, and P. Wang, "Fully Automatic Measurement of Intima-Media Thickness in Ultrasound Images of the Common Carotid Artery Based on Improved Otsu's Method and Adaptive Wind Driven Optimization," *Ultrason. Imaging*, vol. 42, no. 6, pp. 245–260, Nov. 2020, doi: [10.1177/0161734620956897](https://doi.org/10.1177/0161734620956897).
- [8] R. Supriyanti, M. R. Kurniawan, Y. Ramadhani, and H. B. Widodo, "Calculating the area of white spots on the lungs of patients with COVID-19 using the Sauvola thresholding method," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 13, no. 1, p. 315, Feb. 2023, doi: [10.11591/ijece.v13i1.pp315-324](https://doi.org/10.11591/ijece.v13i1.pp315-324).
- [9] M. Rofi'i and D. R. Ningtias, "Local Adaptive Thresholding Menggunakan Metode Sauvola sebagai Tahapan Pra Pengolahan pada Data Citra Isyarat ECG," *Jurnal Teori dan Aplikasi Fisika*, vol. 10, no. 1, p. 103, Jan. 2022, doi: [10.23960/jtaf.v10i1.2904](https://doi.org/10.23960/jtaf.v10i1.2904).
- [10] J. Bhatt, K. A. Hashmi, M. Z. Afzal, and D. Stricker, "A Survey of Graphical Page Object Detection with Deep Neural Networks," *Applied Sciences*, vol. 11, no. 12, p. 5344, Jun. 2021, doi: [10.3390/app11125344](https://doi.org/10.3390/app11125344).
- [11] S. Naik, K. A. Hashmi, A. Pagani, M. Liwicki, D. Stricker, and M. Z. Afzal, "Investigating Attention Mechanism for Page Object Detection in Document Images," *Applied Sciences (Switzerland)*, vol. 12, no. 15, p. 7486, Jul. 2022, doi: [10.3390/app12157486](https://doi.org/10.3390/app12157486).
- [12] S. Satheakeerthy et al., "Zero-Shot LLM-based Visual Acuity Extraction: A Pilot Study," Nov. 08, 2024, doi: [10.21203/rs.3.rs-5287278/v1](https://doi.org/10.21203/rs.3.rs-5287278/v1).
- [13] Z. Yang, S. Zuo, Y. Zhou, J. He, and J. Shi, "A Review of Document Binarization: Main Techniques, New Challenges, and Trends," *Electronics (Basel)*, vol. 13, no. 7, p. 1394, Apr. 2024, doi: [10.3390/electronics13071394](https://doi.org/10.3390/electronics13071394).
- [14] B. Bataineh et al., "A Comprehensive Review on Document Image Binarization," *J. Imaging*, vol. 11, no. 5, 2025, doi: [10.3390/jimaging11050133](https://doi.org/10.3390/jimaging11050133).
- [15] P. R. Sankpal and P. A. Vijaya, "Performance Evaluation of Nested Watermarked Scheme using Objective Image Quality Metrics," *Journal of University of Shanghai for Science and Technology*, vol. 23, no. 06, pp. 306–314, Jun. 2021, doi: [10.51201/JUSST/21/05273](https://doi.org/10.51201/JUSST/21/05273).
- [16] A. W. Dini, S. N. Neyman, and T. Haryanto, "Robust Digital Watermarking on Vital Archives using Hybrid SVD and DWT Methods," *Jurnal RESTI*, vol. 7, no. 5, pp. 1211–1223, Oct. 2023, doi: [10.29207/resti.v7i5.5003](https://doi.org/10.29207/resti.v7i5.5003).
- [17] M. B. DARICI and Z. ERDEM, "A Comparative Study on Denoising from Facial Images Using Convolutional Autoencoder," *Gazi University Journal of Science*, vol. 36, no. 3, pp. 1122–1138, Sep. 2023, doi: [10.35378/gujs.1051655](https://doi.org/10.35378/gujs.1051655).
- [18] N. Hasan, M. S. Islam, W. Chen, M. A. Kabir, and S. Al-Ahmadi, "Encryption Based Image Watermarking Algorithm in 2DWT-DCT Domains," *Sensors*, vol. 21, no. 16, p. 5540, Aug. 2021, doi: [10.3390/s21165540](https://doi.org/10.3390/s21165540).
- [19] I. Hidayati, M. A. Budiman, and M. Zarlis, "Analysis of Embedding Locations in the Subband Frequency DCT on Scanned Images," *Data Science: Journal of Computing and Applied Informatics*, vol. 7, no. 1, pp. 1–14, Jan. 2023, doi: [10.32734/jocai.v7.i1-10359](https://doi.org/10.32734/jocai.v7.i1-10359).
- [20] S. Yang, W. Xiao, M. Zhang, Suhan Guo, J. Zhao, and F. Shen, "Image Data Augmentation for Deep Learning: A Survey," *arXiv Prepr. arXiv2204.08610*, 2022, doi: <https://doi.org/10.48550/arXiv.2204.08610>.
- [21] M. Zhang, Y. Guo, H. Wang, and H. Shangguan, "OFIDA: Object-focused image data augmentation with attention-driven graph convolutional networks," *PLoS One*, vol. 19, no. 5 May, p. e0302124, May 2024, doi: [10.1371/journal.pone.0302124](https://doi.org/10.1371/journal.pone.0302124).
- [22] N. Wulandari, I. Ardiyanto, and H. Adi Nugroho, "A Comparison of Deep Learning Approach for Underwater Object Detection," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 6, no. 2, pp. 252–258, Apr. 2022, doi: [10.29207/resti.v6i2.3931](https://doi.org/10.29207/resti.v6i2.3931).
- [23] B. Chen, C. Yu, S. Zhao, and H. Song, "An Anchor-Free Method Based on Transformers and Adaptive Features for Arbitrarily Oriented Ship Detection in SAR Images," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 17, pp. 2012–2028, 2024, doi: [10.1109/JSTARS.2023.3325573](https://doi.org/10.1109/JSTARS.2023.3325573).

- [24] S. S. Alrumiah, N. Alrebdi, and D. M. Ibrahim, "Augmenting healthy brain magnetic resonance images using generative adversarial networks," *PeerJ Comput. Sci.*, vol. 9, p. e1318, Apr. 2023, doi: [10.7717/peerj-cs.1318](https://doi.org/10.7717/peerj-cs.1318).
- [25] W. Sapitri, Y. N. Kunang, I. Z. Yadi, and M. Mahmud, "The Impact of Data Augmentation Techniques on the Recognition of Script Images in Deep Learning Models," *Jurnal Online Informatika*, vol. 8, no. 2, pp. 169–176, Dec. 2023, doi: [10.15575/join.v8i2.1073](https://doi.org/10.15575/join.v8i2.1073).
- [26] D. Efendi, J. Jasril, S. Sanjaya, F. Syafria, and E. Budianita, "Penerapan Algoritma Convolutional Neural Network Arsitektur ResNet-50 untuk Klasifikasi Citra Daging Sapi dan Babi," *JURIKOM (Jurnal Riset Komputer)*, vol. 9, no. 3, p. 607, Jun. 2022, doi: [10.30865/jurikom.v9i3.4176](https://doi.org/10.30865/jurikom.v9i3.4176).
- [27] M. Rafif, D. U. Kusumaning Putri, and L. Awaludin, "Penggunaan Pre-trained Model untuk Klasifikasi Kualitas Sekrup," *IJEIS (Indonesian Journal of Electronics and Instrumentation Systems)*, vol. 12, no. 2, p. 133, Oct. 2022, doi: [10.22146/ijeis.78112](https://doi.org/10.22146/ijeis.78112).
- [28] C. Anwar, "Deteksi Objek Berbasis Web Menggunakan Tensorflow Js dan Coco Dataset pada Framework React Js," *Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)*, vol. 5, no. 6, pp. 1008–1015, Dec. 2022, doi: [10.32672/jnkti.v5i6.5464](https://doi.org/10.32672/jnkti.v5i6.5464).
- [29] S. Lasniari, J. Jasril, S. Sanjaya, F. Yanto, and M. Affandes, "Klasifikasi Citra Daging Babi dan Daging Sapi Menggunakan Deep Learning Arsitektur ResNet-50 dengan Augmentasi Citra," *Jurnal Sistem Komputer dan Informatika (JSON)*, vol. 3, no. 4, p. 450, Jun. 2022, doi: [10.30865/json.v3i4.4167](https://doi.org/10.30865/json.v3i4.4167).
- [30] A. Wibowo, L. Lusiana, and T. K. Dewi, "Implementasi Algoritma Deep Learning You Only Look Once (YOLOv5) Untuk Deteksi Buah Segar Dan Busuk," *Paspalum: Jurnal Ilmiah Pertanian*, vol. 11, no. 1, p. 123, Mar. 2023, doi: [10.35138/paspalum.v11i1.489](https://doi.org/10.35138/paspalum.v11i1.489).