Bulletin of Electrical Engineering and Informatics
Vol. 15, No. 2, April 2026, pp. 1127~1138
ISSN: 2302-9285, DOI: 10.11591/eei.v15i2.10127

Fine-tuned LayoutLMv3 for Indonesian receipts extraction

Oka Sudana, Ayu Wirdiani, Andre Dwi Winama Putra
Department of Information Technology, Faculty of Engineering, Udayana University, Bali, Indonesia

Article Info

Article history:

Received Feb 21, 2025
Revised Feb 23, 2026
Accepted Mar 5, 2026

Keywords:

Finetuning
Google Vision
LayoutLMv3
Mobile application
Optical character recognition
Receipt extractions

ABSTRACT

Shopping is a transaction that generates a record as a payment receipt. Typically, a receipt is given as a small piece of paper that can be easily lost. It is essential to store the transaction information in the receipt digitally. Keeping the information in a digital form will make it easily accessible and will overcome the problem of easily lost receipts. Currently, the process of transferring receipt information into digital form is still being done manually. Having a system that can extract this information helps speed up the digitalization process tremendously. This research proposes a method that applies finetuning to the LayoutLMv3 model and with the help of optical character recognition (OCR) from Google Vision, can be used to extract transaction information contained in the receipt. The system works by using Google Vision to parse and segment every word contained within the receipt and its bounding box The LayoutLMv3 model will then assign labels to each word, and important words will be extracted. The finetuned LayoutLMv3 model successfully achieved an accuracy of 97.98% on training data and 90% accuracy on real-time test scenarios for extracting information on receipts written in the Indonesian.

Corresponding Author:

Oka Sudana
Department of Information Technology, Faculty of Engineering, Udayana University
Kampus Bukit Jimbaran, Badung, Bali, Indonesia
Email: agungokas@unud.ac.id

1. INTRODUCTION

Nowadays, nearly every transaction generates a payment receipt, whether from shopping at minimarkets, restaurants, malls, hardware stores, or other establishments. These receipts are typically issued as small pieces of paper that contain summaries of transaction details. It is important to keep records in the form of receipts or receipts so that the expenditure of funds can be clearly seen. The best way to store the shopping receipts or receipts is in digital form to ensure that they are not easily lost.

Currently, the extraction process is done manually for each transaction, which takes a lot of time [1]. The existence of a system that can extract information from receipts and save it in a digital format automatically will increase work efficiency. Storing it in digital form also makes it easier to see expenditure information for a certain period. One method that can be used to extract text information is optical character recognition (OCR).

OCR is the process of converting text in an image into machine-readable text format. The image used for text conversion can come from printed text or handwritten characters [2], [3]. OCR technology is a part of artificial intelligence (AI) widely used in automation fields such as document and questionnaire scanning, license plate reading, and document verification, among others [4]. Several model architectures that can be used as a basis for OCR are the long short term memory (LSTM) model and convolutional neural network (CNN). The OCR model is often combined with the natural language processing (NLP) model to improve text reading accuracy [5]. The NLP architecture frequently used in this context is bidirectional

encoder representations from transformers (BERT) [6]. The application of OCR Technology can automatically extract information from receipts and receipts. While some stores currently offer digital receipts, not all do. Furthermore, this application can store expenditure transactions from scanned receipts.

Research into creating an automatic receipt extraction system has been conducted using various methods to detect and localize key information from receipts. Raoui-Outach *et al*. [7] used localization and deep convolution neural networks (DCNN) to segment store signs. Lin *et al*. [8] and Shi *et al*. [9] and used template matching based on prior knowledge and specific signs of a given receipt to locate receipt information. Meng *et al*. [10] used a YOLOv3 model to segment key information on an invoice image with a standardized template. However, some of these systems rely on receipts having a standardized form, which is only sometimes the case, and assume the receipt never changes its template.

Recently, progress in document extraction AI has been significant with the rise of a new topic: Document AI. Document AI is a field of study that aims to provide techniques for understanding, extracting, and classifying documents [11]. Essentially it is an object detection task for document images. An early iteration of the Document AI task is the faster R-CNN model by Schreiber *et al.* [12], which achieved SOTA performance in the ICDAR 2013 Dataset. In recent years progresses on Document AI have been remarkable with the rise of models such as graph convolution network [13] and LayoutLM [14], [15] as well as datasets for benchmarking such as PubLayNet [16] and TableBank [17]. With these new models, we hope to build a system to extract information from a receipt without the need to recognize the exact template of the receipt.

The novelty of this research aims to create a system that can extract information without relying on the receipts template. This research contributes to making it easier for people to store expenditure data from stored receipts. The proposed system uses Google Vision OCR to segment and extract words within the receipt and uses a fine-tuned LayoutLMv3 model to recognize the content of the receipt and extract meaningful information.

2. METHOD

2.1. Finetuning process model LayoutLM

The fine-tuning workflow for LayoutLM, as shown in Figure 1, begins with the collection of an Indonesian Receipt Dataset, followed by preprocessing steps including region of interest (ROI) segmentation, information extraction, and annotation. The training phase involves loading the dataset and extracting both layout and label information from each word in the receipts. This data is input into the LayoutLM Model's autoencoder to ensure consistency with the original pre-training data format. The encoded data is then used to fine-tune the LayoutLM Model. The optimal model identified during training is saved for subsequent receipt detection tasks. System testing is performed by deploying the trained model on a web server. An Android application transmits receipt images to the web server via an application programming interface (API) call, enabling the model to process and extract information from the receipts. System performance is assessed across multiple test scenarios, including various receipt types and differing lighting conditions.

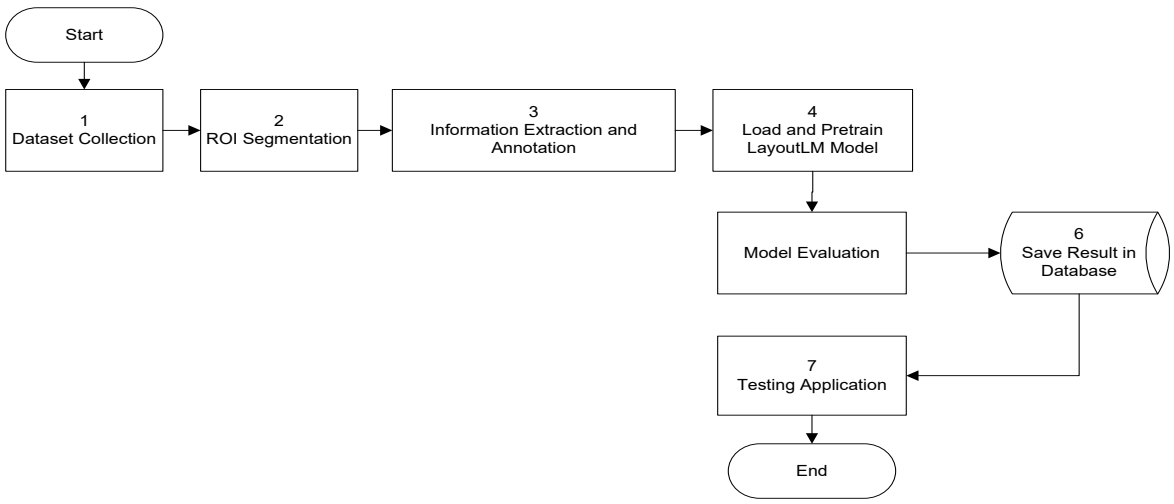


Figure 1. Finetuning process model diagram

2.1. Data collection

The primary dataset was created through the manual collection and photography of 100 shopping receipts from various minimarkets and restaurants. This dataset size was selected because the base LayoutLM Model was fine-tuned using the receipt dataset format adopted in this study to achieve the desired inference results. The collected receipts were subsequently processed for use in model training. The dataset creation process generates a .json file containing information for each word, including its bounding box and the associated label within a receipt.

2.2. Region of interest segmentation

The ROI segmentation process is shown in Figure 2. This segmentation process on receipt images to obtain images that resemble the scan results. This process begins by converting the image to grayscale and applying Gaussian blur to the receipt image. The next step is to apply dilation to the blurred image to prevent the writing on the receipt from being visible as edges to be segmented. The dilated image is then detected using the Canny method to identify the receipt's edges. The Canny edge detection lines are used to separate the receipt image from the background, making the resulting ROI segmentation image resemble a scanned image. The process is shown in Figure 2(a). Figure 2(b) shows an example of the resulting receipt after undergoing ROI segmentation. This image will then be resized to a maximum size of 1000×1000 pixels.

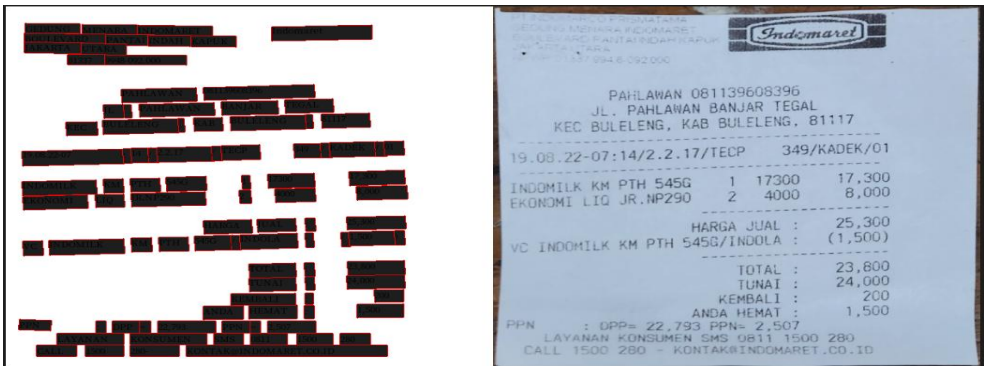
(a) (b)

Figure 2. The ROI segmentation process; (a) data processing: original (top left), grayscale (top right), blur and dilated (bottom left), contour edge (bottom right) and (b) ROI cropping result

2.3. Information extraction and data annotation

The information extraction process involves detecting each word in the receipt using the Layout Parser tool shown in Figure 3. Each detected word and its position are retrieved and temporarily stored before being processed in the annotation process. The annotation process is the process of labeling the primary dataset read by Google Vision. Google Vision API is a machine learning model trained to perform OCR through representational state transfer (REST) and remote procedure call (RPC) APIs. Google Vision API can annotate images and provide labels for each category detected in the image. This annotation process is called automatic image annotation [18]. The automatic image annotation from Google Vision API can extract content from an image to obtain visual information such as labeling images, detecting facial landmarks, and OCR [19]. Data annotation involves the use of the Layout Parser library to extract and segment every word within a receipt using the Google Vision API. The result of information extraction shown in Figure 3(a), where this result is obtained from Figure 3(b) which is the original receipt. The results are then manually annotated by assigning the corresponding labels listed in Table 1.

The information extraction process results in a list of bounding box coordinates and the detected word sequence within the image. These boxes and words then get labeled and saved for finetuning the LayoutLM Model. The product extraction process checks for whether a product name, quantity, and price labels exist in a single line. The products will be extracted if all three are found in a single line. If at least one of the corresponding labels is missing, the system will check for the missing label on the alternate line for receipts that uses multiple lines for each product item.



(a) (b)

Figure 3. Receipt information extraction process; (a) information extraction result and (b) original receipt

Table 1. Receipt annotation code

Label name	Label code
Ignore	0
Store_name_value	1
Date_value	2
Time_value	3
Prod_item_key	4
Prod_item_value	5
Prod_quantity_key	6
Prod_quantity_value	7
Prod_price_key	8
Prod_price_value	9
Subtotal_key	10
Subtotal_value	11
Total_key	12
Total_value	13
Others	14

Table 1 consists of all the labels used in the annotation process. Eight important labels are going to be extracted, which are Store_name_value, Date_value, Time_value, Prod_item_value, Prod_quantity_value, Prod_price_value, Subtotal_value, and Total_value with the other six labels acting as an anchor point to helps determined each important label from the rest. The result of the manually labeled.

Figure 4 shows the results of the annotation process. The figure displays labels for all the words detected Table 1. The fine-tuned model will later be trained to predict labels on various receipts and assign labels accordingly.

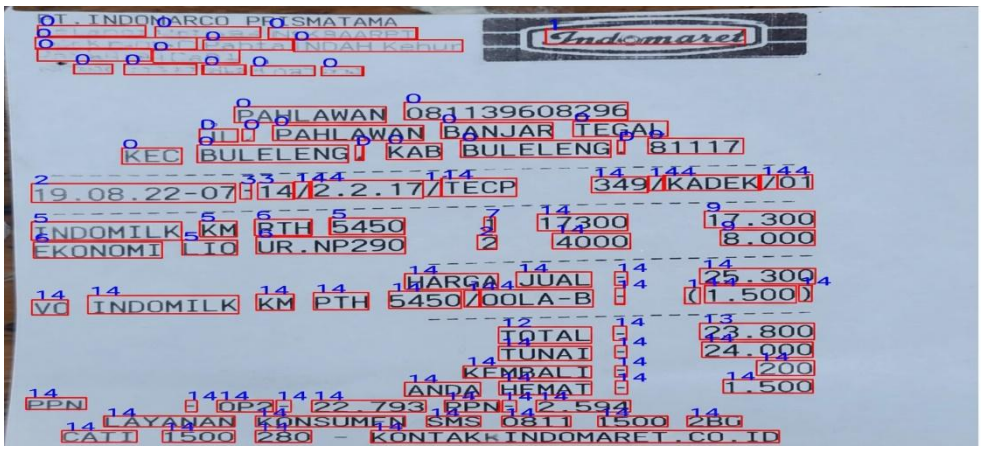


Figure 4. Results of data manually labeled using Table 1 as a reference

2.4. Finetuning LayoutLM Model

The next stage is finetuning of the LayoutLM Model uses the version LayoutLMv3 as the base model. LayoutLM is a document understanding model designed to comprehend the structure of a document. It was developed by considering the developments in NLP, where every NLP model always focuses on text-level manipulation [14], [15]. LayoutLM is created by utilizing the interaction between the text information within a document and its layout using BERT as a reference. This model was developed using data from scanned documents from various categories, such as letters, memos, emails, invoices, news, articles, questionnaires, and resumes [6]. With the recent development of the LayoutLMv3 model aims to analyze visually-rich document understanding (VrDU) where structured information can be automatically extracted. LayoutLMv3 improves upon the original model by integrating masked image modeling (MIM) from bidirectional encoder representation from image Transformers (BEiT) [20] to interpret visual content in the document. Inspired by Vision Transformer (ViT) [21] and vision-and-language Transformer (ViLT) [22], LayoutLMv3 [15] uses linear projection features of image patches before feeding them into the multimodal transformer to remove the need to extract CNN grid features [14], [23] or rely on an object detector like Faster region-based (R-CNN) [24] to extract region features [11], [25]-[27] for image embeddings which require heavy computation bottleneck or region supervision making LayoutLMv3 the first multimodal model in Document AI that does not rely on CNNs to extract image features. The architecture of LayoutLMv3 is shown in Figure 5.

Finetuning is a process similar to transfer learning in convolutional models, where the same model architecture can be used to solve various problems. Finetuning the LayoutLM Model is performed by training the base model with new data from collected receipt. Firstly, the previously gathered data are combined with the wild receipt dataset and split using 80% data for training and 20% for validation, with the total combined data used for training being 1,348, and 492 is used for validation, the detail data is shown in Table 2. The data is then processed using LayoutLMv3 auto processor to obtain the embedding and used for finetuning the base model. The dataset was divided into 80% for training and 20% for testing to ensure that the model had sufficient data to learn representative patterns while still providing an adequate portion of unseen data for an unbiased performance evaluation.

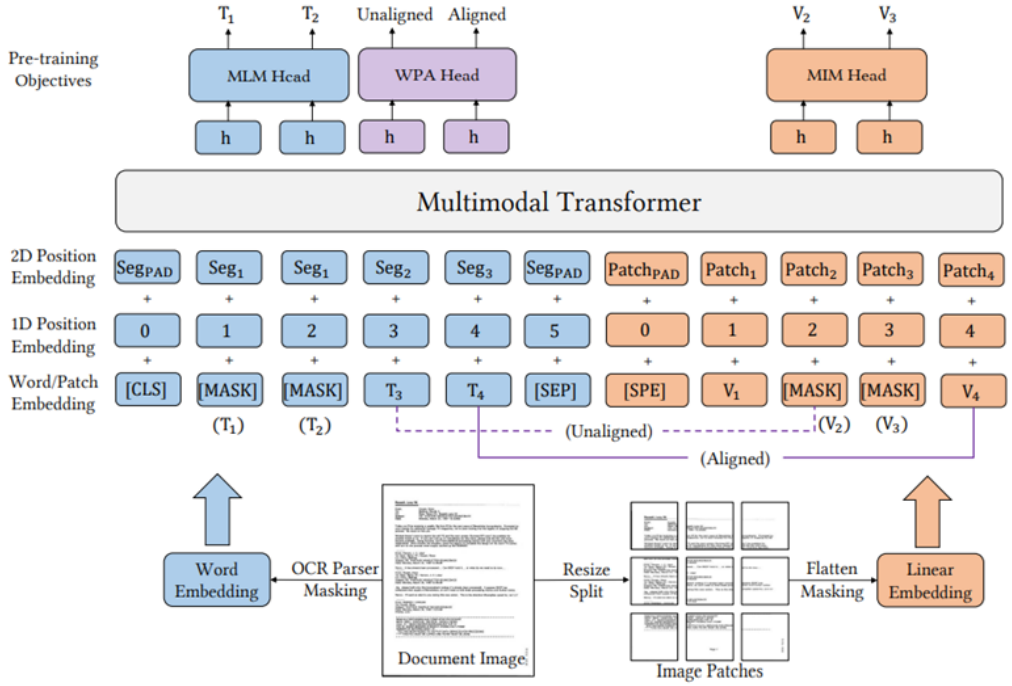


Figure 5. LayoutLMv3 architecture overview [16]

Table 2. Dataset split value

Dataset	Training	Validation
WildReceipt	1268	472
Indonesian receipt (gathered data)	80	20
Total	1348	492

2.5. Model evaluation

The model's performance is evaluated using standard classification evaluation metrics, as shown in the confusion matrix Table 3.

Table 3. Confusion matrix

		Predicted	
Actual		Negative	Positive
	Negative	TN	FP
	Positive	FN	TP

True positive (TP) is a positive label that is successfully predicted as positive, false positive (FP) is a negative label that is incorrectly predicted as positive, false negative (FN) is a positive label that is incorrectly predicted as negative and true negative (TN) is a negative label that is successfully predicted as negative. In this research, the confusion matrix illustrates the detection performance for each label, where correct predictions are represented along the main diagonal. The model is required to distinguish nine labels: Store_name_value, Date_value, Time_value, Prod_item_value, Prod_quantity_value, Prod_price_value, Total_key, and Total_value.

Based on the confusion matrix, accuracy, precision, recall, and F1-score are calculated for each label to assess the model's performance [28]. Precision and recall are other important metrics that provide valuable information regarding how well the model performs [29]. Precision measures how accurate the model is in predicting positive values, while recall measures its strength in predicting positive values.

$$Accuracy = \frac{TP + TN}{TP+FP+FN+TN} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$F - measure = 2 * \frac{precision * recall}{precision + recall} \quad (4)$$

F-measure is calculated using a weighted harmonic mean between precision and recall. F-measure helps to further understand the tradeoff between improving recall and its effect on precision [29].

2.6. System deployment and testing

The system deployment evaluates the model's performance and testing the models in a real-time scenario. Firstly, the finetuned model will be deployed on a web server that takes an image as input from a smartphone. The image will then be processed, and important information will be extracted using the labels assigned by the LayoutLM Model as a reference. Finally, the extracted information is returned to the smartphone, and the results will be evaluated. The overview system shown in Figure 6.

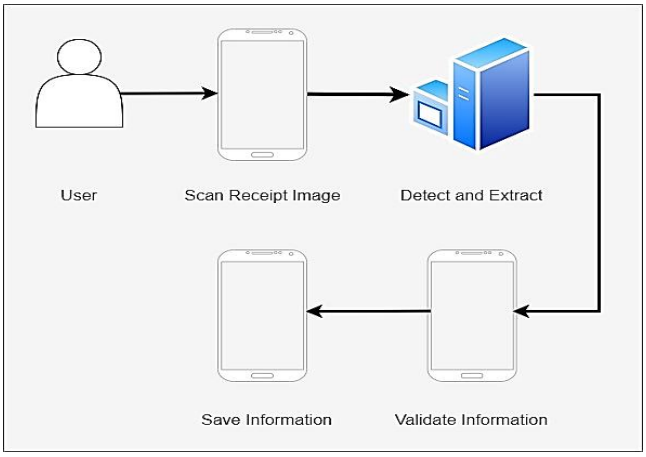


Figure 6. Receipt detection system using mobile

3. RESULTS AND DISCUSSION

3.1. The finetuning model result evaluation

The initial finetuning process on the LayoutLMv3 model shows promising results for classifying important labels on a given receipt. The model is able to differentiate which word is important information and which word should be ignored.

Figures 7 and 8 shows the result of finetuning the LayoutLMv3 base model. The finetuning process on the combined data achieves a training accuracy of 99.2% and an evaluation accuracy of 97.98% with 110 epochs. The publicly available WildReceipt Dataset, which includes receipts written in multiple languages, proves to help improve the model's accuracy for extracting information from Indonesian Receipts compared to using only 100 Indonesian receipts, as shown in Table 4.

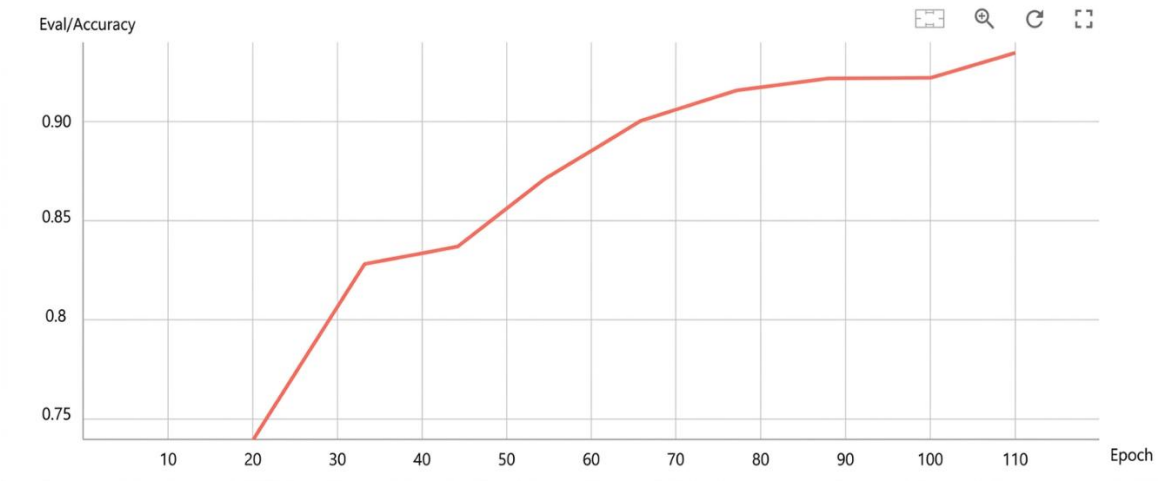


Figure 7. LayoutLMv3 evaluation accuracy with 110 epochs

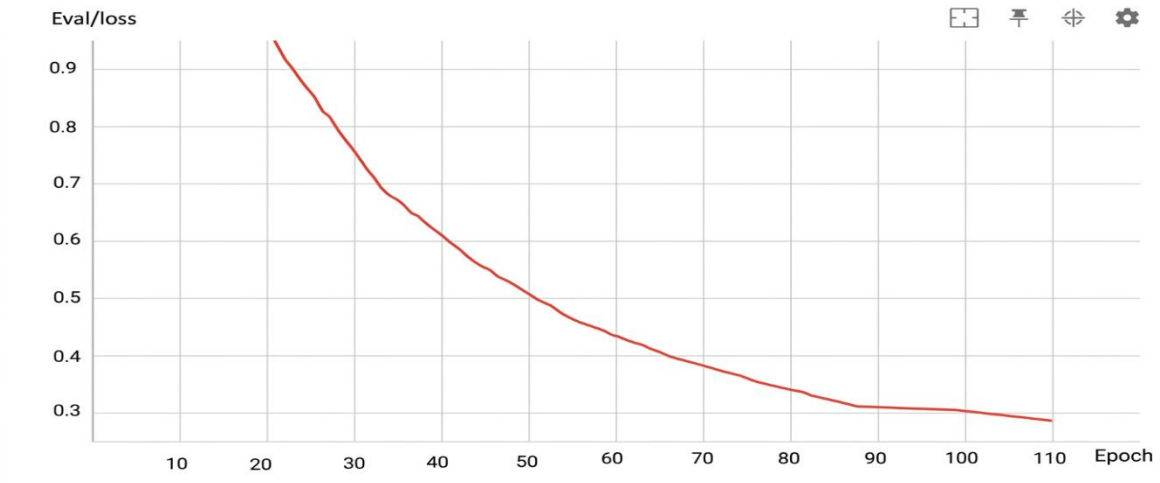


Figure 8. LayoutLMv3 evaluation loss with 110 epochs

Table 4. Training comparison

Model	Precision	Recall	F1-score
Indonesian	0.8889	0.9276	0.9064
Combined	0.9799	0.9798	0.9797

Table 4 compares the result of training with 110 epochs. Based on the results, combined data from 100 gathered Indonesian Receipts combined with the WildReceipt Dataset performed better on all three weighted evaluation categories.

Table 5 presents the model's accuracy in detecting labels under real-time conditions using various receipts captured in optimal lighting environments that shown in Figure 9. The receipts used in this test were obtained from a more diverse set than those used for training and validation. In addition, the test includes the normal condition with receipt templates that the model has never encountered before that shown in Figure 9(a). The Store ID refers to a unique store, while the sample size indicates the number of receipts collected as samples from that store. Several significant issues identified during this test include:

a. The model sometimes incorrectly identifies farewell messages (e.g., "Goodbye," "Thank you") and receipt coupons at the receipt edges as the store name.
b. Part of the receipts identification code is labeled as date information, with a serial number format that resembles date and time information.
c. The model detects total payment twice on some receipts with a non standard total key name such as "Due," "Net Sales," and "Total Sale".

Table 5. Receipt variation test

Store id	Sample size	Average extraction accuracy (%)
1	5	91.7
2	3	93.3
3	3	86.9
4	3	90.9
5	2	100
6	2	94.4
7	4	90.9
8	4	86
9	5	88
10	1	88.3

The results suggest that although the model can accurately detect key information on receipts with moderate spacing between items, it still struggles to extract information from receipts with varying layouts, designs, and fonts. Nevertheless, the proposed system achieves an average extraction accuracy of 90%.

Table 6 shows the model's performance in detecting under different lighting conditions. The result suggests that lighting condition affects the system's segmentation process. Some items cannot be appropriately segmented, which does affect the system's overall accuracy. After further inspection, it is later found that the problem lies in the OCR system rather than the LayoutLM Model. Specifically, the word and bounding box extraction using Vision API sometimes failed to extract information in a dimly lit environment and shaded region. The Figure 9(b) shown the application testing with under different lighting conditions such as normal condition, 30% shade and 80% shade.

Table 6. Various lighting condition test

Store id	Label name	Optimal condition	30% shade	80% shade
1	Store name	Detected	Detected	Detected
	Date	Detected	Detected	Detected
	Time	Detected	Detected	Detected
	Products	Partially detected	Partially detected	Partially detected
	Total	Detected	Detected	Detected
2	Store name	Detected	Detected	Detected
	Date	Detected	Detected	Detected
	Time	Detected	Detected	Detected
	Products	Detected	Detected	Partially detected
	Total	Detected	Detected	Detected
3	Store name	Detected	Detected	Detected
	Date	Not detected	Not detected	Not detected
	Time	Detected	Detected	Detected
	Products	Detected	Detected	Detected
	Total	Not detected	Not detected	Not detected
4	Store name	Detected	Detected	Detected
	Date	Detected	Detected	Detected
	Time	Detected	Detected	Detected
	Products	Detected	Detected	Partially detected
	Total	Detected	Detected	Detected
5	Store name	Not detected	Not detected	Not detected
	Date	Detected	Detected	Detected
	Time	Detected	Detected	Detected
	Products	Detected	Partially detected	Partially detected
	Total	Detected	Detected	Detected

(a)

(b)

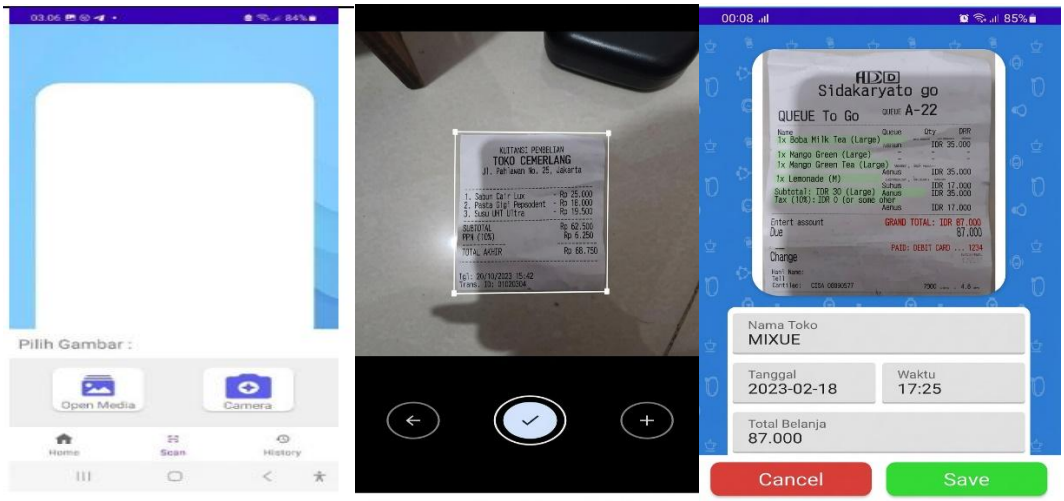
Figure 9. Receipt variation tests sample; (a) normal condition and (b) lighting condition tests sample

3.2. Mobile system testing

The system testing process was conducted using an Android application. The application has several menus that support the inference process on the server receipt shows in Figure 10. The process for receipt inference is as follows. Figure 10(a) shows the scan menu, which contains options for loading images into the system. Selecting the camera menu will launch the camera, which will then forward the results to the cropping process that shows in Figure 10(b). The results then be forwarded to a web server for receipt reading. Figure 10(c) shows the results and information received from reading the receipt via the web server. This display the store name, transaction date, transaction time, total purchase amount, and information about each product purchased. This information can then be reviewed before being saved to the system.

The implications of this research are that the system can read various receipts from minimarkets and restaurants quite well, the system can tolerate crumpled and crossed out receipts when reading receipts quite well, the system provides a cropping feature and selects a gallery that makes users do not need to crop the receipt image with the help of other tools, the system can save the results of reading receipts that can be viewed digitally in the history menu, the storage carried out on the system is in the form of digital information which means that the total shopping within a certain period of time can be seen by applying a filter to the system. The future benefits are for digitizing receipts and recording purchases from receipts, both from supermarkets, minimarkets and restaurants.

A receipt detection system using the LayoutLM Model with Google Vision's OCR successfully extracted information from receipt without needing to recognize the template of each receipt. The application of the LayoutLM Model can replace the rough estimation process used in Lin's automatic receipt recognition system research [8]. Based on the tests conducted, the model's accuracy during the initial evaluation, which was 97.98%, fluctuated during real-time testing. The model still failed to perform optimally on several receipt variations with closely spaced information and too small fonts. These variations caused the system's test accuracy to drop to 90%.



(a) (b) (c)

Figure 10. Receipt mobile system application; (a) image source menu view, (b) cropping process, and (c) preview display of receipt reading results

4. CONCLUSION

The fine-tuning process applied to LayoutLMv3 results in enhanced efficacy when it comes to spotting and deciphering receipts written in Indonesian. The system underwent training utilizing gathered Indonesian Receipts alongside the readily accessible WildReceipt Dataset, which encompasses receipts composed in a variety of different languages. Using the combined dataset proves to be significant in increasing the accuracy of the model for extracting information from Indonesian Receipts compared to using 100 Indonesian Receipts. The best-performing model is able to achieve an accuracy of 97.98% for predicting keywords on Indonesian Receipts after being trained for 110 epochs. Despite the promising results, there are limitations that affect the system's accuracy. The problem found during the system testing is that the system has not been able to detect receipt optimally in dark conditions or covered by shadows. This problem is caused by the OCR extraction process through Google Vision, which sometimes fails to extract important information from notes. As future work, this research aims to facilitate receipt information extraction, enabling users to more easily split bills when making payments at restaurants or in other transactional scenarios.

FUNDING INFORMATION

This research was funded by the Institute for Research and Community Service, Udayana University with contract number: B/229.361/UN14.4.A/PT.01.03/2025.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Oka Sudana	✓	✓		✓	✓	✓			✓	✓	✓	✓	✓	✓
Ayu Wirdiani	✓	✓		✓	✓	✓		✓	✓	✓	✓	✓		
Andre Dwi Winama Putra	✓		✓	✓		✓	✓	✓	✓	✓				

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

Data availability is not applicable to this paper as no new data were created or analyzed in this study.

REFERENCES

[1] F. Kosadi, W. Ginting, and V. Merliana, "Digital Receipts of Online Transactions in the Reconciliation Process and the Preparation of Financial Reports," *Journal of Indonesian Economy and Business*, vol. 36, no. 1, pp. 31–50, 2021, doi: 10.22146/jieb.59884.

[2] 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. 142642–142668, 2020, doi: 10.1109/ACCESS.2020.3012542.

[3] V. Kumar, P. Kaware, P. Singh, R. Sonkusare, and S. Kumar, "Extraction of Information from Bill Receipts using Optical Character Recognition," in *2020 International Conference on Smart Electronics and Communication (ICOSEC)*, 2020, pp. 72–77, doi: 10.1109/ICOSEC49089.2020.9215246.

[4] A. Qaroush, A. Awad, M. Modallal, and M. Ziq, "Segmentation-based, Omnifont Printed Arabic Character Recognition without Font Identification," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 6, pp. 3025–3039, Jun. 2022, doi: 10.1016/j.jksuci.2020.10.001.

[5] M. Hajiali, J. R. F. Cacho, and K. Taghva, "Generating Correction Candidates for OCR Errors using BERT Language Model and FastText SubWord Embeddings," *Lecture Receipts in Networks and Systems*, vol. 283, pp. 1045–1053, 2022, doi: 10.1007/978-3-030-80119-9_69.

[6] J. Devlin, M.-W. Chang, K. Lee, K. T. Google, and A. I. Language, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019, pp. 4171-4186, doi: 10.18653/v1/N19-1423.

[7] R. Raoui-Outach, C. Million-Rousseau, A. Benoit and P. Lambert, "Deep Learning for Automatic Sale Receipt Understanding," in *2017 Seventh International Conference on Image Processing Theory, Tools and Applications (IPTA)*, Montreal, QC, Canada, 2017, pp. 1-6, doi: 10.1109/IPTA.2017.8310088.

[8] C.-J. Lin, Y.-C. Liu, and C.-L. Lee, "Automatic Receipt Recognition System Based on Artificial Intelligence Technology," *Applied Sciences*, vol. 12, p. 853, 2022, doi: 10.3390/app12020853.

[9] S. Shi, C. Cui, and Y. Xiao, "An Invoice Recognition System using Deep Learning," in *2020 International Conference on Intelligent Computing, Automation and Systems (ICICAS)*, 2020, pp. 416–423, doi: 10.1109/ICICAS51530.2020.00093.

[10] Y. Meng, R. Wang, J. Wang, J. Yang, and G. Gui, "IRIS: Smart Phone Aided Intelligent Reimbursement System using Deep Learning," *IEEE Access*, vol. 7, pp. 165635–165645, 2019, doi: 10.1109/ACCESS.2019.2953501.

[11] Y. Xu, M. Li, L. Cui, S. Huang, F. Wei, and M. Zhou, "LayoutLM: Pre-training of Text and Layout for Document Image Understanding," *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Aug. 2020, vol. 20, pp. 1192-1200, doi: 10.1145/3394486.3403172.

[12] S. Schreiber, S. Agne, I. Wolf, A. Dengel, and S. Ahmed, "DeepDeSRT: Deep Learning for Detection and Structure Recognition of Tables in Document Images," in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, 2017, pp. 1162–1167, doi: 10.1109/ICDAR.2017.192.

[13] X. Liu, F. Gao, Q. Zhang, and H. Zhao, "Graph Convolution for Multimodal Information Extraction from Visually Rich Documents," *in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers),* 2019, pp. 32–39, doi: 10.18653/v1/N19-2005

[14] Y. Xu *et al.*, "LayoutLMv2 : Multi-modal Pre-training for Visually-rich Document Understanding," in *KDD '20: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 1192–1200, doi: 10.1145/3394486.3403172.

[15] Y. Huang, T. Lv, L. Cui, Y. Lu, and F. Wei, "LayoutLMv3: Pre-training for Document AI with Unified Text and Image Masking;" in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 4083-4091, doi: 10.1145/3503161.3548112.

[16] X. Zhong, J. Tang, and A. J. Yepes, "PubLayNet: Largest Dataset Ever for Document Layout Analysis," in *2019 International Conference on Document Analysis and Recognition (ICDAR)*, Sydney, NSW, Australia, 2019, pp. 1015-1022, doi: 10.1109/ICDAR.2019.00166.

[17] M. Li, L. Cui, S. Huang, F. Wei, M. Zhou, and Z. Li, "TableBank: A Benchmark Dataset for Table Detection and Recognition," *arXiv*, doi: 10.48550/arXiv.1903.01949.

[18] P. Baker and L. Collins, "Creating and analysing a multimodal corpus of news texts with Google Cloud Vision's Automatic Image Tagger," *Applied Corpus Linguistics*, vol. 3, no. 1, Apr. 2023, doi: 10.1016/j.acorp.2023.100043.

[19] K. Saputra, D. Rahmaastri, K. Setiawan, D. Suryani, and Y. Purnama, "Mobile Financial Management Application using Google Cloud Vision API," *Procedia Computer Science*, vol. 157, pp. 596–604, Feb. 2019, doi: 10.1016/j.procs.2019.09.019.

[20] H. Bao, L. Dong, S. Piao, and F. Wei, "BEiT: BERT Pre-Training of Image Transformers," *The Eleventh International Conference on Learning Representations (ICLR 2022)*, 2022.

[21] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL),* 2021, pp. 2579–2591.

[22] W. Kim, B. Son, and I. Kim, "ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision," *in Proceedings of the 38th International Conference on Machine Learning (ICML)*, vol. 139, 2021, pp. 5583–5594.

[23] S. Appalaraju, B. Jasani, B. U. Kota, Y. Xie, and R. Manmatha, "DocFormer: End-to-End Transformer for Document Understanding," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 978-983, doi: 10.1109/ICCV48922.2021.00103.

[24] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 39, no. 6, pp. 1137–1149*,* Jun. 2017, doi: 10.1109/TPAMI.2016.2577031.

[25] J. Gu *et al.*, "Unified Pretraining Framework for Document Understanding," in *Proceedings of 35th Conference on Neural Information Processing Systems (NeurIPS 2021)*, 2021, pp. 39-50, doi: 10.5555/3540261.3540265.
[26] P. Li *et al.*, "SelfDoc: Self-Supervised Document Representation Learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* 2021, pp. 5652–5660, doi: 10.1109/CVPR46437.2021.00560.
[27] R. Powalski, Ł. Borchmann, D. Jurkiewicz, T. Dwojak, M. Pietruszka, and G. Pałka, "Going Full-TILT Boogie on Document Understanding with Text-Image-Layout Transformer," in *International Conference on Document Analysis and Recognition,* Feb. 2021, pp. 732-747, doi: 10.1007/978-3-030-86331-9_47.
[28] A. Kulkarni, D. Chong, and F. A. Batarseh, "5 - Foundations of Data Imbalance and Solutions for a Data Democracy," *Data Democracy*, 2020, pp. 83–106, doi: 10.1016/B978-0-12-818366-3.00005-8.
[29] H. Singh, *Practical Machine Learning and Image Processing: For Facial Recognition, Object Detection, and Pattern Recognition Using Python*, 2019, doi: 10.1007/978-1-4842-4149-3.

BIOGRAPHIES OF AUTHORS

Oka Sudana received the graduate from Department of Informatics from Institute Technology of Sepuluh Nopember Surabaya (ITS) in 1997. He received the Master degree in Department of Electrical Engineering from Faculty of Engineering Gadjah Mada University (UGM) Yogyakarta, in 2001. He received the Doctoral degree in Department of Doctoral Program in Engineering Science at Udayana University Bali, in 2020. Also, he received the Engineer degree in Professional Engineer Program at Udayana University Bali, in 2024. Currently, he is Research Imaging System Laboratory in Department of Information Technology. Faculty of Engineering, Udayana University Bali Indonesia. His research interests include image processing and pattern recognition, information technology implementation in computer vision and culture. He can be contacted at email: agungokas@unud.ac.id.

Ayu Wirdiani received the graduate from Department of Electrical Engineering from Faculty of Engineering, Udayana University in 2003. She received the Master degree in Department of Electrical Engineering from Faculty of Engineering Udayana University Bali, in 2011. She received the Doctoral degree in Department of Doctoral Program in Engineering Science at Udayana University Bali, in 2024. Also, she received the Engineer degree in Professional Engineer Program at Udayana University Bali, in 2024. Currently, she is Research Imaging System Laboratory in Department of Information Technology. Faculty of Engineering, Udayana University Bali Indonesia. Her research interests include image processing and pattern recognition, information technology implementation in computer vision and culture. She can be contacted at email: ayuwirdiani@unud.ac.id.

Andre Dwi Winama Putra received the graduate from Department of Information Technology from Faculty of Engineering, Udayana University in 2024. His research interests include image processing, information technology implementation in Computer Vision. Now he worked as a Data Analyst in Software Company. He can be contacted at email: andre002wp@gmail.com.