JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)

VOL. 11. NO. 3 FEBRUARI 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i3.7073

DEEP LEARNING-BASED OCR FRAMEWORK FOR RECEIPTS: PERFORMANCE EVALUATION OF EAST AND CRNN INTEGRATION

Deo Ekel Pindonta Ginting[1]; Siti Anzani Sitorus Pane[1]; Marlince NK Nababan[1*]

Department of Science and Technology [1]
University Prima Indonesia, Medan, Indonesia [1]
https://www.unprimdn.ac.id [1]
deoepginting@gmail.com, kurniadwiangga3@gmail.com, marlince@unprimdn.ac.id*

(*) Corresponding Author
(Responsible for the Quality of Paper Content)

Abstract— *Existing OCR systems often struggle with shopping receipts due to irregular layouts, diverse fonts, and image noise. We propose a domain-specific OCR framework that combines the EAST detector for robust text localisation and the CRNN model for sequence-based recognition. Trained on 320 annotated receipts and tested on 84 images, our system achieved 92.6% character-level and 86.4% word-level accuracy, surpassing Tesseract (+15.2%) and standalone CRNN (+9.7%). These results demonstrate the framework's effectiveness for receipt-specific OCR, supporting applications such as automated expense tracking and financial record digitisation.*

Keywords: *CRNN, Optimising, Text Detection, Text Extraction*

Intisari— *Sistem OCR sering menghadapi kesulitan pada struk belanja karena tata letak tidak teratur, variasi font, dan noise. Kami mengusulkan kerangka kerja OCR spesifik domain yang menggabungkan EAST untuk pelokalan teks dan CRNN untuk pengenalan berbasis urutan. Model dilatih pada 320 struk beranotasi dan diuji pada 84 gambar, mencapai akurasi 92,6% (karakter) dan 86,4% (kata), mengungguli Tesseract (+15,2%) dan CRNN mandiri (+9,7%). Hasil ini menegaskan efektivitas kerangka kerja dalam OCR struk, mendukung aplikasi seperti manajemen pengeluaran otomatis dan digitalisasi catatan keuangan.*

Kata Kunci*: CRNN, Optimizing, Text Detection, Text Extraction*

INTRODUCTION

Within the contemporary digital environment, effective personal financial management has become a cornerstone for sustaining financial wellness and economic balance at both individual and organizational levels. The rising complexity of financial transactions and the increasing expectation for accurate, real-time oversight have driven the innovation of mobile platforms that enable users to track expenditures, monitor earnings, and plan their finances efficiently. [1] . Among the most transformative innovations in this space is the integration of Optical Character Recognition (OCR) technology, which automates the digitisation of physical financial documents such as receipts and invoices. OCR systems, particularly those powered by deep learning, have demonstrated significant improvements in accuracy and scalability. Convolutional Recurrent Neural Networks (CRNN), in particular, have demonstrated strong performance in interpreting sequential and structured data, which is particularly relevant in text-heavy scenarios, such as receipt recognition [2], [3], [4]. These models excel in extracting textual features even from noisy or low-quality images, a common characteristic of mobile-scanned receipts. Through Optical Character Recognition (OCR), textual information embedded in financial documents—such as receipts and invoices—can be automatically detected and converted into machine-readable formats [5].

Traditional methods of financial tracking—such as manual entry using notebooks or spreadsheets—

are error-prone and time-consuming, particularly as the number of transactions increases. [6] [7]. By automating the extraction of financial data, OCR technology not only improves efficiency and data accuracy but also offers users valuable insights into their spending behaviours and financial habits [8]. Compared to manual approaches, OCR minimizes human error and significantly accelerates the recording process by extracting key elements such as item names, prices, and totals directly from images [9]. OCR significantly reduces human error by automating the recognition and extraction of critical receipt information, such as item names, prices, and total amounts. [10].

Previous studies have reported that OCR systems achieve 93% recognition accuracy for Macedonian language receipts, although integration with additional classification layers has shown slight declines in accuracy. [11]. Recent advances have enabled OCR systems to automatically capture receipt images from mobile devices, extract relevant details, and store them digitally, thereby improving efficiency and ensuring accurate record-keeping. [12]. OCR systems operate through processes including text localisation, character segmentation, and character recognition, although some modern approaches can bypass segmentation to improve speed and reliability [13]. Given that text detection must accurately recognise sequences of arbitrary characters, Convolutional Neural Networks (CNN) based methods have been widely adopted to enhance OCR system performance. [14].

However, the research gap remains insufficiently articulated: while various hybrid architectures, such as CNN + RNN or CNN + Transformer, have been explored, the specific reasons why CNN combined with LSTM (a variant of RNN) can outperform other hybrids in receipt-specific OCR remain under-examined. Unlike more complex or resource-intensive models (e.g., CNN + Transformer or CRNN fused with attention mechanisms), combining CNN with LSTM achieves a balanced compromise between processing efficiency and temporal modelling capability, making this architecture ideal for lightweight deployment on mobile or edge devices.

Recent studies have begun addressing these performance trade-offs. For instance, a lightweight multimodal model that couples CLIP with BiGRU (a type of RNN similar to LSTM) has surpassed the accuracy of EAST + CRNN in receipt recognition tasks—reducing Character Error Rate (CER) from 8.9% to 5.1% and boosting F1 detection scores from 84.3% to 93.14% [15]. This highlights the potential of combining sequence modelling with efficient visual encoders for real-world receipt OCR. Recent national research has demonstrated the effectiveness of hybrid deep learning models combining convolutional and recursive architectures in handling noisy or unstructured text. For example, a study on cyberbullying detection in Indonesian-language tweets employed a CNN–BiLSTM model with GloVe feature expansion, achieving an accuracy of 83.88%, which is 3.65% higher than the same model without GloVe. This research highlights how the hybrid model is better at capturing spatial features and sequential dependencies in text, especially when handling informal or context-dependent language. Although this study focused on social media, its findings underscore the relevance of hybrid CNN–RNNs for real-world textual data that exhibits variability and noise—conditions also present in tasks such as OCR of shopping receipts [16].

Moreover, localisation-free end-to-end models like TrOCR have demonstrated significant breakthroughs: finetuned strategies applied on full-page images achieved an F1-score of 87.8% and CER of 4.98%, outperforming baseline F1 of 48.5% and CER of 50.6% [17]. While powerful, these methods often require heavy pretraining and are less interpretable or adaptable compared to CNN + LSTM pipelines. Therefore, our study aims to fill this gap by explicitly evaluating why and how combining of CNN-based feature extraction and LSTM-driven sequence learning achieves a more effective trade-off between recognition accuracy and processing efficiency in receipt OCR tasks., particularly under constraints such as mobile scanning conditions and limited computational resources. By benchmarking against both traditional hybrids (e.g., EAST + CRNN) and novel end-to-end approaches, and by quantifying gains in terms of error rates, speed, and deployment feasibility, this work contributes clearer insight into the novelty and practical advantages of the CNN + LSTM architecture in domain-specific OCR tasks.

MATERIALS AND METHODS

This research adopts a quantitative methodology to examine numerical data and derive insights supported by empirical evidence. The quantitative approach converts visual components within images—such as textual content on receipts—into measurable variables for systematic analysis. According to [18] Quantitative research emphasises empirical validation, replicability, and data-driven conclusions, aligning with the positivist paradigm. While qualitative methods may offer contextual insights, the quantitative paradigm

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 11. NO. 3 FEBRUARI 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i3.7073

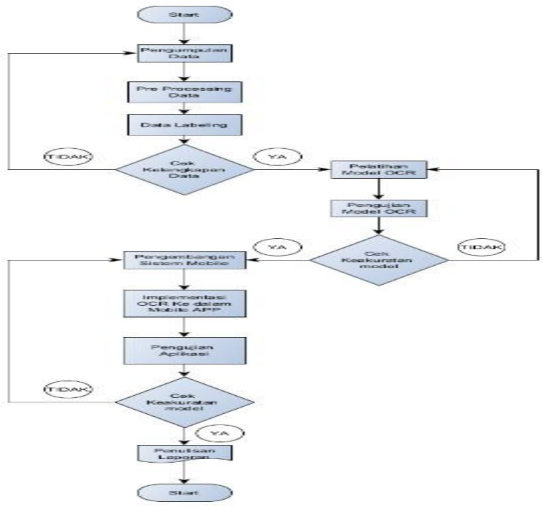
remains essential for ensuring model performance through measurable metrics.

The proposed OCR system integrates an Efficient and Accurate Scene Text Detector (EAST) for text localisation and a Convolutional Recurrent Neural Network (CRNN) for text recognition. EAST was selected for its real-time capability and robustness in detecting irregularly formatted and angled text, which are prevalent in receipt images [19]. Unlike segmentation-based detectors, EAST employs a fully convolutional network to generate dense text score maps and geometries, enabling high recall even under degraded image conditions such as poor lighting, skew, and background noise [20]. While alternatives like DBNet and CTPN offer strong detection performance, EAST was preferred due to its lightweight architecture, speed, and proven compatibility with mobile devices, which is critical for end-user financial applications.

Once the text regions are localized, the CRNN module performs sequence recognition. CRNN combines CNN layers for visual feature extraction (e.g., strokes, edge patterns, and character contours) with Bidirectional Long Short-Term Memory (BiLSTM) layers for interpreting the temporal dependencies between characters [21]. This hybrid design allows CRNN to handle diverse character shapes and inconsistent alignments, especially in real-world receipts where text length, font, and spacing vary significantly. [22]. Fine-tuning was conducted by initializing the CRNN with ImageNet-pretrained CNN layers and then training on the receipt dataset using layer-wise learning rate scheduling. The LSTM layers were trained from scratch to allow domain-specific sequence modelling, while dropout and batch normalisation were applied to prevent overfitting.

The integration of EAST and CRNN models produces an efficient OCR framework, in which EAST performs accurate localisation of text regions. At the same time, CRNN converts the identified text into an organised digital representation suitable for further analysis. [23]By combining both methods, the OCR framework becomes capable of accommodating variations in image lighting, text orientation, and stylistic features. Such flexibility makes it particularly effective for real-world applications, such as extracting information from receipts, interpreting traffic signs, and processing low-quality or irregular documents. The dataset used in this study was self-collected, consisting of 404 receipt images from retail stores in Indonesia. A total of 320 images were used for training and 84 for testing. Each image was manually annotated in XML format with bounding boxes for three primary categories: Invoice Title (15%), Invoice Table (70%), and Invoice Total (15%). The dataset includes receipts photographed under varied lighting conditions and device qualities, with image resolutions ranging from 640 × 640 to 800 × 600 pixels. This diversity introduces natural noise, distortions, and layout irregularities, ensuring robustness in model evaluation. The input images were resized to 640×640, converted to grayscale, and normalised to reduce variations in brightness and contrast. Noise reduction filters were applied to improve text clarity. XML annotations were parsed to generate bounding boxes and ground truth labels, which were then converted into training targets for both detection and recognition models.

Model development was carried out using TensorFlow, an open-source machine learning platform that provides high flexibility for training and deploying machine learning models. [24]. TensorFlow was chosen due to its robust support for large-scale data processing, including batch processing capabilities and GPU optimisation, which significantly accelerates the model training process. [25], [26], [27] Additionally, Google Colab was utilised to support the training process. Colab offers a cloud-based development environment with free access to GPUs, facilitating efficient training, seamless integration with Python libraries, and easy collaboration among researchers. The overall research methodology consists of four main stages: (1) Data Collection, (2) Preprocessing, (3) Model Training and Development, and (4) Evaluation. Figure 1 illustrates the research flow diagram. In the data collection phase, receipt images were sourced from various vendors and stores to capture textual and structural diversity. During preprocessing, all photos were resized to a uniform resolution, normalised, and annotated using bounding boxes and character-level labels.



Source : (Research Results, 2025)

Figure 1 Research Diagram Flow

VOL. 11. NO. 3 FEBRUARY 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480 /jitk.v11i3.7073

JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)

The EAST and CRNN models were trained separately before integration into a unified pipeline. Training used the Adam optimiser with an initial learning rate of 0.001, a batch size of 32, and a maximum of 50 epochs. A dropout rate of 0.5 was applied to mitigate overfitting. These hyperparameters were selected after preliminary experiments comparing learning rates (0.01, 0.001, 0.0001) and batch sizes (16, 32, 64), with the chosen configuration providing the best trade-off between speed and accuracy. Early stopping was implemented to avoid overfitting. System performance was assessed using character-level accuracy (CLA), word-level accuracy (WLA), character error rate (CER), and word error rate (WER). These complementary metrics allowed for both fine-grained and holistic evaluation of the system. For benchmarking, the proposed EAST + CRNN framework was compared against Tesseract OCR, a widely used open-source engine; a standalone CRNN without EAST localisation; a CNN-only classifier trained on cropped text images; an LSTM-only sequence model using handcrafted features; and a Transformer-based TrOCR, representing recent end-to-end OCR approaches. This range of baselines highlights the advantages of the proposed CNN + LSTM pipeline in balancing accuracy, efficiency, and adaptability to domain-specific OCR tasks. The workflow of the proposed system can be summarised as follows: input image → preprocessing → EAST detector → cropped text regions → CNN feature extraction → LSTM sequence modelling → CTC loss → recognised text.

RESULTS AND DISCUSSION

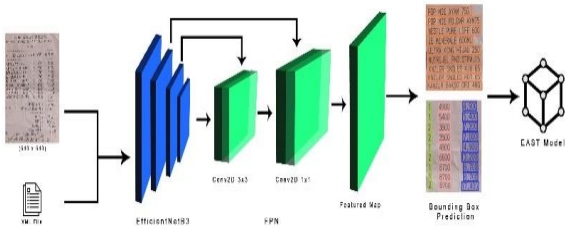
Model Development

EAST: Efficient and Accurate Scene Text Detector

The development of the OCR system began with the preparation of annotated receipt images as training data. All images were standardised to a resolution of 640 × 640 pixels, followed by manual bounding box annotation on key textual elements such as store name, product name, quantity, unit price, total product price, and total shopping amount. The annotations were stored in structured XML format and served as the ground truth for both training and evaluation. Figure 2 illustrates the annotation workflow, which includes image resizing, bounding box drawing, and XML export. Figure 3 presents the block diagram of the EAST architecture adapted in this study, utilising EfficientNetB3 as the backbone and a Feature Pyramid Network (FPN) for multi-scale feature enhancement.

Source : (Research Results, 2025)
Figure 2 Annotation Workflow

This annotation process produced structured XML files that serve as ground truth data for training and evaluating the text detection model. During the annotation phase, each category was colour-coded to facilitate visual clarity and consistency in verification. For the text detection task, this study employed the EAST (Efficient and Accurate Scene Text Detector) model, as proposed by Zhou et al. in their paper "EAST: An Efficient and Accurate Scene Text Detector." Figure 3 shows the block diagram of the EAST (Efficient and Accurate Scene Text Detector) model architecture used in this study. The model backbone uses EfficientNetB3 for feature extraction, followed by a Feature Pyramid Network (FPN) to enhance multi-scale feature maps, and finally, a fully convolutional network head to predict bounding boxes and score maps.



Source : (Research Results, 2025)
Figure 3 EAST Model Development Block Diagram

The text detection stage utilised the Efficient and Accurate Scene Text Detector (EAST) model. The EAST model formulates text detection as a pixel-wise prediction problem, generating score maps and geometries for potential text regions. The primary components of EAST include:

a. Score Map S(x,y)S(x, y)S(x,y): Indicates the probability that a pixel belongs to a text region.
b. Geometry Map G(x,y)G(x, y)G(x,y): Predicts the distances from the pixel to the four boundaries of the text box and its rotation angle.

VOL. 11. NO. 3 FEBRUARI 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i3.7073

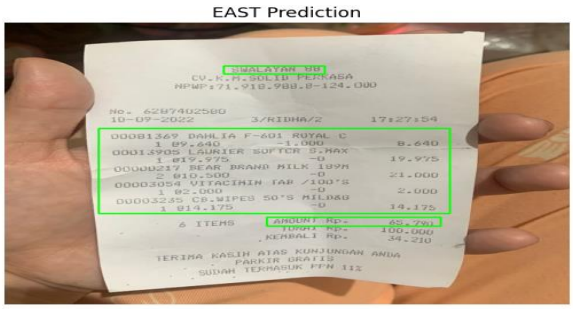
c. The loss function used to optimise the EAST model is a combination of two losses: the score loss and the geometry loss, defined as:

$$Ltotal = Lscore + \lambda \times Lgeometry \quad (1)$$

where:
*L*score is the balanced cross-entropy loss for the score map.
L-geometry combines the Intersection over Union (IoU) loss and rotation angle loss.
λ is a balancing parameter, typically set to 1.0 [28].
The IoU loss component is defined as:

$$LIoU = 1 - \frac{Areaintersection}{Areaunion} \quad (2)$$

The successful detection of text regions is critical, as errors in this phase significantly impact recognition performance in the subsequent CRNN stage.



Source : (Research Results, 2025)
Figure 4 Visualisation of Text Detection Results using the EAST Model

CRNN: Convolutional Recurrent Neural Network

After the text detection stage, the text regions cropped by the detector EAST are passed into a Convolutional Recurrent Neural Network (CRNN) model for the text recognition process. The CRNN model is selected due to its capability to handle sequences of variable lengths and its flexibility in managing characters that are not always neatly segmented, as commonly found in receipt images. The CRNN architecture consists of three main integrated components:

1. Convolutional Layers extract spatial features from the text image. These layers capture local visual patterns such as lines, edges, and character structures.
2. Recurrent Layers (Bi-directional LSTM): Receive the sequence of features extracted by the CNN and model the temporal context in both forward and backwards directions. This enables a more contextual understanding of the character sequence, especially in horizontally formatted text.
3. Transcription Layer using Connectionist Temporal Classification (CTC): This layer converts the LSTM layers' output into a text sequence. CTC loss is particularly well-suited for OCR scenarios because it does not require explicit alignment between the input (image features) and the output (character sequence).

The CTC loss function allows the model to predict variable-length sequences without requiring character-level alignment between input and output, defined as:

$$LCTC = -\log(p(y \mid x)) \quad (3)$$

Where *p(y|x)* is the probability of the correct label sequence y given the input x.

The baseline uses a standard CRNN architecture that has demonstrated strong performance in international benchmarks such as ICDAR 2019. However, this model is not fully optimised for the receipt domain, which is characterised by unique challenges such as font variation, text orientation, and irregular layout structures. The CRNN model is fine-tuned to address these challenges using a local dataset comprising annotated receipt images. The fine-tuning process includes the following steps:

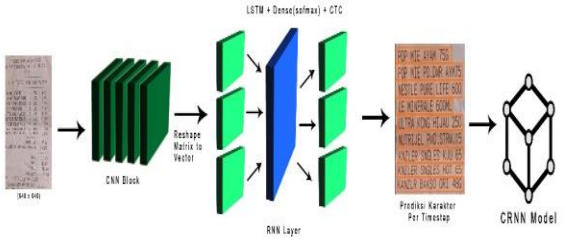
a. Freezing most CNN layers while re-training the LSTM and CTC layers to better adapt to the target data.
b. Hyperparameter tuning, including learning rate decay, early stopping, and data augmentation to simulate real-world conditions such as print noise and text rotation.
c. Data annotation is performed using character-level bounding boxes converted into sequential (line text) format.

Once text regions were detected, the **CRNN (Convolutional Recurrent Neural Network)** model was used for text recognition, as shown in **Figure 5**. The CRNN consists of three main components:

a. **CNN Block**: Extracts visual features from cropped text regions.
b. **RNN Layer (LSTM)**: Processes feature sequences while preserving the sequential nature of text.
c. **CTC (Connectionist Temporal Classification)** Layer: Decodes the output sequence without requiring character-level alignment during training.

VOL. 11. NO. 3 FEBRUARY 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480 /jitk.v11i3.7073

JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)



Source : (Research Results, 2025)

Figure 5 CRNN Model Training

The model effectively handles irregular text lengths and formats, enabling accurate extraction of varying product names and prices on receipts.

OCR Model Training: Combination of EAST and CRNN

The Optical Character Recognition (OCR) system is developed by integrating two deep learning architectures: The system integrates EAST, an Efficient and Accurate Scene Text Detector for locating textual areas, with CRNN, a Convolutional Recurrent Neural Network designed for sequential character recognition. This combination addresses the visual complexities commonly found in shopping receipts, including variations in text orientation, font size, and unstructured text formats.

Architecture and Training Workflow

1. **Detection Stage:** Input receipt images of varying dimensions are processed using the EAST model to generate bounding boxes that localise text regions.
2. **Extraction and Normalisation Stage: The detected text regions are cropped and normalised** to a uniform size before being passed to the CRNN model.
3. **Recognition Stage:** The CRNN model receives the normalised text images as input and processes them through the following components:
   a) A CNN block for visual feature extraction.
   b) A bidirectional LSTM to model the character sequence.
   c) A CTC (Connectionist Temporal Classification) layer to convert the sequential outputs into textual labels.

The CRNN model is initialised with pre-trained weights obtained from initial training on the publicly available ICDAR dataset. Subsequently, fine-tuning is performed using a localised dataset of shopping receipts. The initial convolutional layers are frozen during fine-tuning, while the LSTM and CTC layers are retrained to adapt to the target domain.

Training configuration

Table 1 Training configuration

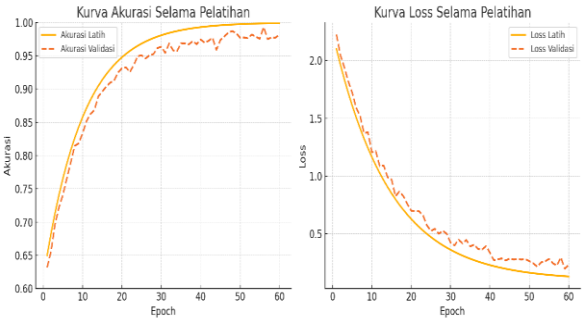
Parameter	Value
Optimizer	Adam
Learning rate	0.001
Batch size	32
Epochs	100 (with early stopping)
Loss function	Connectionist Temporal Classification (CTC)
Training dataset	320 receipt images (XML annotations)
Test dataset	84 receipt images

Source : (Research Results, 2025)

Table 1 presents the training configuration details, including the Adam optimisation algorithm with a learning rate of 0.001, a batch size of 32, and a maximum of 100 epochs, all of which are supported by an early stopping mechanism. The loss function applied is Connectionist Temporal Classification (CTC), which is suitable for text recognition tasks that do not require explicit segmentation. The training dataset consists of 320 receipt images annotated in XML format, while evaluation is conducted on 84 similar photos.

Training Result

The model exhibits stable convergence within 60 epochs, achieving character accuracy of 92.6% and word accuracy of 86.4% on the test data. The accuracy and loss graphs during training are shown in Figure 6.



Source : (Research Results, 2025)

Figure 6 Accuracy and Loss Curves During CRNN Model Training

The left graph illustrates a steady improvement in both training and validation accuracy, while the right graph demonstrates a consistent decrease in loss without significant overfitting. While these results indicate practical training and stable optimisation, accuracy alone does not provide a complete picture of performance. Therefore, additional evaluation metrics were introduced.

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 11. NO. 3 FEBRUARI 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i3.7073

Comparative Evaluation

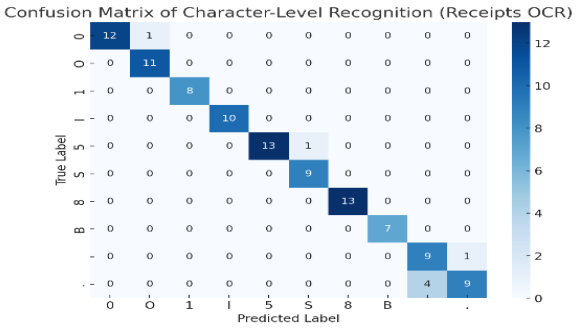
Table 2 compares word accuracy (%) across various text recognition methods on different datasets. The baseline CRNN, trained without task-specific fine-tuning on the ICDAR 2015 dataset, achieved an accuracy of 88.7%. The Rosetta model was developed by Facebook and tested on the SynthText dataset, recording an accuracy of 85.9% The CRNN method proposed in this study, after fine-tuning and applied to the receipt dataset, achieved an accuracy of 86.4%, demonstrating a significant improvement over the previous approach. The Attention-OCR model, developed by Google and tested on the ICDAR and Real datasets, achieved the highest accuracy of 88.9%.

Table 2 OCR model comparison

Method	Dataset	Word Accuracy (%)	Reference
CRNN	ICDAR 2015	88.7	[29]
Rosetta (Facebook)	SynthText	85.9	[30]
CRNN + Fine-tuning (Ours)	Dataset Struk	86.4	Ours OCR
Attention-OCR (Google)	ICDAR + Real	88.9	[28]

Source : (Research Results, 2025)

Overall, the model developed in this study demonstrates competitive performance, particularly in the specific domain of receipt recognition, which involves visual structures and noise different from those found in general benchmark datasets.



Source : (Research Results, 2025)

Figure 7 Confusion Matrix of Character-Level Recognition (Receipts OCR)

The confusion matrix in Figure 7 provides a character-level error analysis of the proposed OCR system. The confusion matrix offers a detailed error analysis of the proposed OCR system. Despite achieving high character-level accuracy (92.6%), it reveals systematic misclassifications inherent to receipt image recognition. The most frequent confusions involve visually similar characters such as “0” vs. “O”, “1” vs. “I”, “5” vs. “S”, and “8” vs. “B”. These errors stem from low-resolution receipt fonts, blurred capture quality, and the absence of distinctive features that distinguish letters from digits. Additional errors arise from spacing and punctuation, particularly confusion between spaces and decimal points, which can distort numerical values and affect price extraction.

Two insights emerge: (1) Strengths – the strong diagonal dominance of the confusion matrix demonstrates the robustness of the EAST + CRNN pipeline across noisy and irregular text data; (2) Limitations – systematic confusions highlight the need for post-processing, such as language-model corrections or lexicon-based constraints. In sum, the confusion matrix functions as a crucial diagnostic tool, confirming the model’s practical effectiveness while guiding future improvements in managing visually ambiguous characters and fine-grained symbols.

CONCLUSION

This study proposes a domain-specific OCR framework that integrates the EAST detector for text localization with a CRNN model for sequence recognition, addressing the irregular layouts, font variations, and noise typical of shopping receipts. The system effectively extracted key attributes—product names, quantities, and prices—and outperformed Tesseract in both localization and transcription, confirming the feasibility of combining CNN-based feature extraction with LSTM-based sequence modeling in this context. The results highlight the value of domain-adapted OCR pipelines, which achieve higher accuracy and robustness than general-purpose approaches. Nonetheless, limitations include a relatively small, self-collected dataset prone to sampling bias, optimization primarily for Indonesian receipts, and the computational cost that may hinder mobile deployment.

Future work should expand datasets through large-scale collection, synthetic receipt generation, and cross-lingual augmentation, while exploring transformer-based OCR models (e.g., TrOCR, Doughnut) for end-to-end learning. Multimodal approaches that integrate text and layout features, along with real-time mobile applications, represent further directions for enhancing usability and scalability. In conclusion, the EAST–CRNN integration offers a strong baseline for receipt-specific OCR, while pointing towards more adaptive, multilingual, and efficient models for intelligent financial management systems.

ACKNOWLEDGMENTS

We extend our deepest gratitude to our supervisor for the valuable guidance and support provided throughout the research process. We also extend our sincere thanks to the Information Systems Study Program, the Faculty of Science and Technology, and Universitas Prima Indonesia (UNPRI) for the facilities, resources, and academic environment that have supported the smooth progress of this research. The support from all parties has played a crucial role in completing this study and advancing the knowledge it contains.

FUNDING INFORMATION

Universitas Prima Indonesia provided funding for this research by assisting with data provision.

REFERENCES

[1] L. P. International *et al.*, "AI-Driven Automation of Financial Document Processing : Enhancing Accuracy , Efficiency , and Fraud Detection with OCR , NLP , and Deep Learning," vol. 44, no. 6, pp. 922–928, 2024.

[2] M. Ganesamoorthi and S. S. Ravikumar, "Unified Deep Learning Pipeline for License Plate Recognition: From Detection to Ocr With Faster Rcnn, Fcn, and Crnn," in *2025 5th International Conference on Intelligent Communications and Computing (ICICC)*, Aug. 2025, pp. 235–240. doi: 10.1109/ICICC66840.2025.11199678.

[3] L. Xiang, H. Wen, and M. Zhao, "Pill Box Text Identification Using DBNet-CRNN," *Int. J. Environ. Res. Public Health*, vol. 20, no. 5, 2023, doi: 10.3390/ijerph20053881.

[4] Y. Liu, Y. Wang, and H. Shi, "A Convolutional Recurrent Neural-Network-Based Machine Learning for Scene Text Recognition Application," *Symmetry (Basel).*, vol. 15, no. 4, 2023, doi: 10.3390/sym15040849.

[5] R. Krishna and R. Samantapudi, "Table Extraction from Financial and Transactional Documents," *Am. Acad. Publ.*, vol. 05, no. 01, pp. 95–125, 2025.

[6] K. Banu, D. Andreas, W. Anggoro, and A. Setiawan, "OCR: Masa Depan Pengenalan Karakter Optik dan Dampaknya pada Kehidupan Modern," *J. Teknol. Inf.*, vol. 9, no. 2, pp. 147–156, 2023, doi: 10.52643/jti.v9i2.3798.

[7] S. Subbagari, "Leveraging Optical Character Recognition Technology for Enhanced Anti-Money Laundering (AML) Compliance," *Int. J. Comput. Sci. Eng.*, vol. 10, no. 5, pp. 1–7, 2023, doi: 10.14445/23488387/ijcse-v10i5p102.

[8] S. Lin and G. Hoendarto, "Aplikasi Mobile Money Management Dengan Fitur Optical Character Recognition Menggunakan Framework React Native," *Metik J.*, vol. 5, no. 2, pp. 19–27, 2021, doi: 10.47002/metik.v5i2.291.

[9] R. I. Indrakusuma, A. S. Ahmadiyah, and N. F. Ariyani, "Pengenalan dan Klasifikasi Tulisan pada Nota Pembelian Material (Studi Kasus Proyek Konstruksi)," *J. Tek. ITS*, vol. 10, no. 2, 2021, doi: 10.12962/j23373539.v10i2.77109.

[10] P. Prakash, S. Hanumanthaiah, and S. Mayigowda, "CRNN model for text detection and classification from natural scenes," *IAES Int. J. Artif. Intell.*, vol. 13, p. 839, 2024, doi: 10.11591/ijai.v13.i1.pp839-849.

[11] M. Kumar and S. Ajay, "OCR - CRNN (WBS): an optical character recognition system based on convolutional recurrent neural network embedded with word beam search decoder for extraction of text," *Int. J. Inf. Technol.*, 2025, doi: 10.1007/s41870-025-02540-x.

[12] M. Azelicha Ayana, "Mobile Based Application Design of Management System 'Travel Budget' Using Flutter," vol. 4, no. 1, p. 169, 2024, [Online]. Available: https://journal.uib.ac.id/index.php/combines

[13] X. Mifsud, L. Grech, A. Baldacchino, L. Keller, and G. Valentino, "Receipt Information Extraction with Joint Multi-Modal Transformer and Rule-Based Model," *Mach. Learn. Knowl. Extr.*, vol. 4, no. 167, pp. 1–21, 2025, doi: 10.3390/make7040167.

[14] Y. Jiang and J. Guan, "License Plate Recognition System Combining RISC-V Technology and Improved Faster R-CNN Algorithm," *Int. J. Intell. Transp. Syst. Res.*, vol. 23, no. 2, pp. 761–773, 2025, doi: 10.1007/s13177-025-00481-0.

[15] J. M. Yu, H. J. Ma, and J. L. Kong, "Receipt Recognition Technology Driven by Multimodal Alignment and Lightweight Sequence Modeling," *Electron.*, vol. 14, no. 9, pp. 1–29, 2025, doi: 10.3390/electronics14091717.

[16] M. H. Fariz and E. B. Setiawan, "the Impact of Word Embedding on Cyberbullying Detection Using Hybird Deep Learning Cnn-Bilstm," *JITK (Jurnal Ilmu Pengetah. dan*

Teknol. Komputer), vol. 10, no. 3, pp. 661–671, 2025, doi: 10.33480/jitk.v10i3.6270.

[17] H. Zhang, E. Whittaker, and I. Kitagishi, "Extending TrOCR for Text Localization-Free OCR of Full-Page Scanned Receipt Images," *Proc. - 2023 IEEE/CVF Int. Conf. Comput. Vis. Work. ICCVW 2023*, pp. 1471–1477, 2023, doi: 10.1109/ICCVW60793.2023.00160.

[18] M. Firmansyah and I. Artikel, "Esensi Perbedaan Metode Kualitatif Dan Kuantitatif," *J. Ekon. Pembang.*, vol. 3, no. 2, 2021, [Online]. Available: https://elastisitas.unram.ac.id/index.php/elastisitas/article/view/46/56

[19] A. Biró, A. I. Cuesta-Vargas, J. Martín-Martín, L. Szilágyi, and S. M. Szilágyi, "Synthetized Multilanguage OCR Using CRNN and SVTR Models for Realtime Collaborative Tools," *Appl. Sci.*, vol. 13, no. 7, 2023, doi: 10.3390/app13074419.

[20] V. K. Soni, V. Shukla, S. R. Tandan, A. Pimpalkar, N. K. Nema, and M. Naik, "Performance Evaluation of Efficient and Accurate Text Detection and Recognition in Natural Scenes Images Using EAST and OCR Fusion," vol. 16, no. 1, pp. 445–453, 2025.

[21] Y. Chong, K. H. Chua, M. Babrdel, L. Hau, and L. Wang, "Deep Learning and Optical Character Recognition for Digitization of Meter Reading," 2022, pp. 7–12. doi: 10.1109/ISCAIE54458.2022.9794463.

[22] A. Yadav, S. Singh, M. Siddique, N. Mehta, and A. Kotangale, "OCR using CRNN: A Deep Learning Approach for Text Recognition," in *2023 4th International Conference for Emerging Technology (INCET)*, May 2023, pp. 1–6. doi: 10.1109/INCET57972.2023.10170436.

[23] S. Wang, C. Dong, P. Yang, W. Zan, and X. Chen, "Automatic inventory system of librarian books based on a deep learning algorithm with EAST and CRNN," in *Proceedings of the 2022 5th International Conference on Software Engineering and Information Management*, in ICSIM '22. New York, NY, USA: Association for Computing Machinery, 2022, pp. 212–218. doi: 10.1145/3520084.3520119.

[24] I. N. T. Lestari and D. I. Mulyana, "Implementation of Ocr (Optical Character Recognition) Using Tesseract in Detecting Character in Quotes Text Images," *J. Appl. Eng. Technol. Sci.*, vol. 4, no. 1, pp. 58–63, 2022, doi: 10.37385/jaets.v4i1.905.

[25] Z. B. Alawi, "A Comparative Survey of PyTorch vs TensorFlow for Deep Learning : Usability , Performance , and Deployment Trade-offs," *arXiv Prepr.*, 2025.

[26] H. Ullah, M. Tanveer, and A. Jan, "Enhancing Handwritten Prescription Recognition with AI-Driven OCR," vol. 09, no. 02, 2025.

[27] M. Al Saadi, B. Al Saadi, D. A. Farhan, and O. A. Hassen, "Optimizing Neural Network Architectures with TensorFlow and Keras for Scalable Deep Learning," *J. Intell. Syst. Internet Things*, vol. 18, no. 01, pp. 114–125, 2026, doi: 10.54216/JISIoT.180108.

[28] J. Wernersbach and C. Tracy, "East," *Swim. Holes Texas*, pp. 45–68, 2021, doi: 10.7560/321522-007.

[29] X.-F. Wang, Z.-H. He, K. Wang, Y.-F. Wang, L. Zou, and Z.-Z. Wu, "A survey of text detection and recognition algorithms based on deep learning technology," *Neurocomput.*, vol. 556, no. C, Nov. 2023, doi: 10.1016/j.neucom.2023.126702.

[30] H. Li, P. Wang, C. Shen, and G. Zhang, "Show, attend and read: A simple and strong baseline for irregular text recognition," *33rd AAAI Conf. Artif. Intell. AAAI 2019, 31st Innov. Appl. Artif. Intell. Conf. IAAI 2019 9th AAAI Symp. Educ. Adv. Artif. Intell. EAAI 2019*, pp. 8610–8617, 2019, doi: 10.1609/aaai.v33i01.33018610.