AgroExpense-OCR: Vision Transformer-Based Optical Character Recognition for Agricultural Receipt Digitization

Wongpanya S. Nuankaew [1], Chanapa Phikhason [1], Watthana Kayowaen [1], Tatsaneewan Yenwattana [1], and Pratya Nuankaew [2,*]

[1] Department of Computer Science, School of Information and Communication Technology, University of Phayao, Phayao 56000, Thailand

[2] Department of Digital Business, School of Information and Communication Technology, University of Phayao, Phayao 56000, Thailand

Email: wongpanya.nu@up.ac.th (W.S.N.); 65020834@up.ac.th (C.P.); 65022128@up.ac.th (W.K.); 65024568@up.ac.th (T.Y.); pratya.nu@up.ac.th (P.N.)

*Corresponding author

***Abstract*—This study concentrates on the development of AgroExpense-OCR, an Optical Character Recognition (OCR) system for agricultural expense receipts, employing the Vision Transformer (ViT) architecture to improve the recognition of both printed and handwritten text. Utilizing a dataset of over 15,000 systematically prepared items, experimental results demonstrate that FastViT achieves the fastest convergence and lowest Character Error Rate (CER), despite transient instability during handwritten sample training. The best character-level performance is achieved by FastViT, with a test loss of 0.03144 and Certificate of Analysis (CER) of 0.0086 using greedy decoding and 0.0091 with beam search. Evaluation results demonstrated superior performance, with printed receipts attaining an accuracy of 96.80% and handwritten receipts 92.40%, surpassing traditional CNN-RNN methodologies. The AgroExpense-OCR application was created for both mobile and web platforms, including features such as receipt scanning, automatic expense categorization, cost analysis, and budget notifications. User satisfaction assessments indicated a high level of acceptance, thereby confirming the practical applicability of the system. This research makes a significant contribution to the advancement of digital agriculture and establishes a foundation for integration with future smart farm management systems.**

***Keywords*—agricultural receipts, digital agriculture, optical character recognition, smart farming, vision transformer**

I. Introduction

Efficient financial recordkeeping is essential to modern agricultural management. Farmers and agribusinesses handle numerous transactions involving inputs such as seeds, fertilizers, and pesticides, usually documented through paper receipts [1]. However, manual entry and storage of such data are time-consuming and error-prone, especially for smallholder farmers who often lack access to digital infrastructure. Automating receipt data extraction is therefore essential to improve accuracy and support informed decision-making. As agriculture undergoes digital transformation, machine learning-based automation has become crucial for enhancing productivity and financial transparency [2].

Optical Character Recognition (OCR) is an artificial intelligence and machine learning-based technology that converts textual content from images or documents into structured digital data. It supports both printed and handwritten text across multiple languages, making it widely applicable in industries such as finance, healthcare, and document management. Despite its versatility, OCR still faces challenges in recognizing handwriting, mathematical expressions, and complex layouts due to variations in writing styles and symbol ambiguity [3]. Conventional OCR models typically rely on Convolutional Neural Networks (CNNs) and Convolutional Recurrent Neural Networks (CRNNs). For instance, Lin *et al.* [1] combined a Connectionist Text Proposal Network (CTPN) with a CRNN to extract itemized receipt data [1], A CNN-based OCR model for handwritten characters and digits, demonstrating strong performance on constrained datasets, was proposed by Raza *et al.* [4]. However, such CNN-based approaches struggle with multilingual and unstructured handwritten text. Similarly, Anakpluek *et al.* [5] reported that even optimized Tesseract OCR models still face limitations when processing Thai documents, emphasizing the need for domain-adapted OCR frameworks.

Recent advances in Vision Transformers (ViTs) have greatly enhanced OCR performance by capturing global contextual relationships within document images through self-attention mechanisms. Models such as No-Recurrence sequence-to-sequence Text Recognizer (NRTR), Sequential Transformation Attention-Based Network (STAN) model [6], and DOTA demonstrate strong scalability for text recognition across complex layouts and

Manuscript received September 14, 2025; revised December 21, 2025; accepted January 7, 2026; published May 13, 2026.

degraded imagery [7]. Luan *et al.* [8] proposed a lightweight scene text recognition model (LSTR) based on the ViT-based scene text recognition model. LSTR uses only 7 million parameters yet achieves a good balance of accuracy and speed, surpassing previous approaches. Nuankaew *et al.* [9] extended ViT's use to Thai handwriting in agricultural accounting, showing it outperforms CNNs. These studies highlight ViT's versatility for semi-structured text tasks.

Thailand presents a particularly relevant context for OCR research in agriculture, where smallholder farmers remain heavily reliant on paper-based documentation for recording transactions. Despite initiatives to promote digital agriculture, handwritten and partially printed receipts are still widely used. Thai script, especially in handwritten form, poses significant challenges for OCR systems due to the absence of clear word boundaries, complex glyph structures and diacritics, and high variability in individual handwriting styles [10]. Chamchong *et al.* [11] achieved only moderate accuracy on the BEST2019 dataset, while Nimsuk *et al.* [12] improved performance using multiple deep networks for offline Thai handwriting recognition. For Thai numerals, Khunratchasana and Treenuntharath [13] demonstrated the effectiveness of a CNN-based model for recognizing Thai numerals on a custom dataset. However, in real-world agricultural settings, complexity increases due to low-quality images, uneven lighting, crumpled paper, and the presence of tabular or domain-specific symbols on receipts. A more recent study by Nuankaew *et al*. [9] further demonstrated that Vision Transformer-based models significantly improve Thai handwriting recognition for agricultural bookkeeping, reinforcing the importance of Thai-specific datasets and domain-adapted OCR frameworks for enhancing document digitization accuracy.

Receipts in structured domains such as restaurants are typically generated from predefined menus or databases before printing, resulting in standardized item names and layouts that require minimal manual input. In contrast, agricultural receipts vary greatly in both format and linguistic clarity. When handwritten, item descriptions are often abbreviated or shortened to fit limited space, particularly when the handwriting is large or uneven. Longer product names or technical details may be reduced for readability, sometimes causing incomplete or ambiguous information. Even printed agricultural receipts frequently contain domain-specific terminology, irregular formatting, which complicates text extraction and interpretation.

Building upon the findings of Basir *et al.* [10], which underscore the need for interoperable, accurate, and automated data systems in farming, this research extends these concepts to address the challenges of document digitization in Thailand's agricultural sector. Despite significant progress in OCR and Vision Transformer-based text recognition, existing systems still struggle with multilingual agricultural documents that include both Thai and English content, handwritten entries, linguistic complexity, and poor image quality. To bridge these gaps, this research proposes AgroExpense-OCR, a Vision Transformer-based OCR framework designed to accurately extract and digitize textual information from agricultural receipts. The system aims to enhance data accuracy, reduce manual entry errors, and support Thailand's transition toward intelligent and data-driven agricultural management.

The research objectives are as follows:

- To design and develop a Vision Transformer-based OCR model capable of recognizing and classifying textual information from agricultural receipts, including both printed and handwritten data, with a focus on accuracy and handling semi-structured documents.
- To implement the AgroExpense-OCR application for mobile and web platforms that enable automatic recording, categorization, and analysis of agricultural expenses extracted from receipts.
- To evaluate user satisfaction among farmers and stakeholders regarding the AgroExpense-OCR application in terms of usability, recognition accuracy, and its effectiveness in agricultural cost management.

This research aims to integrate a Vision Transformer-based OCR system for digitizing agricultural receipts, representing a major step toward accurate and scalable expense management in agriculture. By tackling both the technical challenges of semi-structured document recognition and the linguistic complexity of Thai script, the proposed AgroExpense-OCR application bridges computer vision technology with real agricultural applications. Combining deep learning, mobile implementation, and user-focused evaluation, the system improves accounting efficiency and supports data-driven, sustainable farm management.

II. Literature Review

Recent advances in OCR have rapidly progressed through advancements in deep learning and computer vision, enabling accurate extraction of both printed and handwritten text across multiple languages and document formats. Modern OCR models, including convolutional and transformer-based architectures, have enhanced accuracy, flexibility, and processing efficiency. This section provides an overview of major OCR approaches, their fundamental principles, and current limitations in handling semi-structured and multilingual documents, particularly those containing complex scripts such as Thai.

In receipt digitization, ViTs are promising because they can capture long-range dependencies in document layouts [14]. For example, Atienza [15] introduced ViTSTR, a scene text recognition model based on a vision transformer, which achieved competitive accuracy on standard benchmarks while being significantly faster and more parameter-efficient than CNN-RNN baselines. Similarly, Li *et al.* [16] proposed TrOCR, an end-to-end OCR model that employs a Transformer for image encoding and another for text decoding. TrOCR was pretrained on large synthetic datasets and fine-tuned on real-world text images, resulting in state-of-the-art performance on printed, handwritten, and scene text

datasets. These studies show the effectiveness of pure transformer architectures for OCR tasks, bypassing the need for recurrent networks or language-model post-processing.

Applying ViT-based OCR specifically to receipts has started to draw attention. Yu *et al.* [14] observed that transformer-based models, including ViTs, "have been explored for receipt recognition tasks due to their ability to capture long-range dependencies". In practice, digital receipt processing faces unique challenges, such as complex layouts, low-quality scans, and mixed content. New multimodal models, like CLIP or LayoutLM variants, combine visual and textual cues to improve field extraction; however, they are often too large for on-device use. Therefore, lightweight ViT solutions are highly sought after. Additionally, analysis beyond basic OCR is required: datasets like the SROIE receipts (ICDAR 2019) and the recent CORU benchmark [17] encompass tasks from text localization to semantic parsing on receipts. For example, the CORU dataset includes thousands of annotated receipts with key fields, such as merchant, date, total, and items, to support detailed analysis [17]. These resources help develop end-to-end pipelines that not only recognize text with ViT-based models but also automatically categorize receipts and generate insights.

Classification and automated analysis necessary for effective receipt digitization extend beyond mere Optical Character Recognition (OCR). The process involves the classification and organization of receipts, as well as downstream data analysis. Recent research indicates that integrating vision-based classification methods with transformer-based text understanding can significantly improve document sorting capabilities. Dutta *et al.* [18] introduced VisFormers, a hybrid model engineered for complex document classification: a pretrained convolutional neural network (e.g., VGG) initially categorizes the document image into a broad class, which is subsequently refined by feeding OCR output into a Transformer for detailed text-based classification. This two-stage framework exemplifies how vision transformers can complement traditional CNN pipelines by incorporating textual context from OCR, thereby enabling robust categorization of documents and receipts.

For automated archiving and expense analysis, such frameworks enable agricultural receipts to be automatically categorized, for example, as "fertilizer," "equipment," and stored in relevant databases. Moreover, the detailed annotations in modern receipt datasets facilitate in-depth analysis, such as expense summarization or item-level classification. The CORU dataset [17], for example, includes item-specific annotations that allow models to classify purchased products or detect anomalies. Overall, integrating ViT-based OCR with downstream Transformer classifiers supports both the digitization of receipts and the automation of expense analytics. As research advances the development of ViT architectures for OCR and document understanding [18, 19], solutions like AgroExpense-OCR can achieve high accuracy in reading receipts while enabling advanced data categorization and insights extraction.

Geng *et al.* [20] proposed LW-ViT, a lightweight Vision Transformer designed for offline handwritten Chinese character recognition. The model simplifies MobileViT's architecture by reducing both Transformer blocks and MV2 layers, resulting in substantially fewer parameters and lower computational demands while maintaining a high classification accuracy of about 95.8%. Despite these advantages, the evaluation was limited to clean, pre-segmented character images, which fail to capture real-world complexities such as overlapping strokes, uneven spacing, and visual noise. Furthermore, the model was not assessed as part of a complete OCR system that includes character detection and segmentation. Therefore, although LW-ViT performs efficiently and accurately in controlled settings, its effectiveness in practical handwriting recognition remains uncertain. Within the same period, Geng et al. developed LSTR, a Vision Transformer-based lightweight model designed to enhance scene text recognition by addressing attention drift and reducing computational demands. The architecture incorporates a position-enhancement branch to improve the alignment between positional and visual features, along with a visual-enhancement module that strengthens spatial representation within the encoder-decoder framework. Experimental results on both synthetic and real datasets indicate that LSTR achieves higher recognition accuracy than baseline and other lightweight models like CRNN, ViTSTR, and MGP, while maintaining efficient inference performance. Although the enhanced decoder slightly increases computational complexity, it contributes to notable gains in accuracy. The ablation experiments further validate the effectiveness of the proposed modules in mitigating attention drift. Nonetheless, the evaluation was conducted under controlled conditions, leaving its robustness in real-world scenarios and applicability within complete OCR systems for future investigation [8].

While Nuankaew *et al.* [9] demonstrated the strong potential of Vision Transformers (ViT) for Thai handwritten character recognition, achieving over 95% accuracy on a structured dataset collected from diverse age groups. Their model effectively distinguished Thai consonants, vowels, tones, and numerals under controlled conditions, providing a solid proof of concept. However, the system was trained and tested only on clean, segmented characters, which did not reflect the complexity of real-world handwriting that often includes noise, overlapping strokes, and inconsistent spacing. The study did not evaluate the model in a full end-to-end pipeline that includes character detection and segmentation from raw, unstructured documents, such as agricultural receipts. As a result, while the classifier performed well in isolation, its practical effectiveness in real-world applications remained unproven.

The literature review emphasizes the increasing importance of Vision Transformers (ViTs) in OCR, especially for digitizing agricultural receipts. Recent studies like ViTSTR, LW-ViT, and TrOCR show that ViTs surpass traditional CNN-RNN models in recognizing

both printed and handwritten text, while also better managing the structural complexity of receipts. However, there are still few applications specifically focused on agricultural receipts. Projects like the CORU dataset help deepen understanding, allowing for automated classification, structured data storage, and expense analysis. Overall, ViTs are not only improving OCR accuracy but are also becoming the foundation of systems like AgroExpense-OCR, which aim to support cost analysis and efficient management of agricultural expenses in the digital age.

III. Materials

The Materials and Methods section of this study outline the procedures and tools used to develop and assess the AgroExpense-OCR system, which employs Vision Transformer (ViT) architecture as the main OCR engine for recognizing text from agricultural receipts, including both printed and handwritten documents. The research process was carefully designed to cover all stages, such as sample selection, data collection and preparation, OCR model and application development, expense categorization and analysis, and system validation with actual farmers. This approach ensures that the methodology is well-organized, transparent, reproducible, and capable of producing results that demonstrate practical usefulness in agriculture (see Fig. 1).

A. Population and Sample

The study involved 80 participants, including 27 farmers, 10 agricultural cost stakeholders, 13 IT specialists, and 30 computer science students from the Department of Computer Science, School of Information and Communication Technology, University of Phayao. Their families are engaged in agriculture or farm production. The participants were divided into two main groups: the first group provided data for developing a manual analysis model. In contrast, the second group evaluated user satisfaction with the AgroExpense-OCR app.

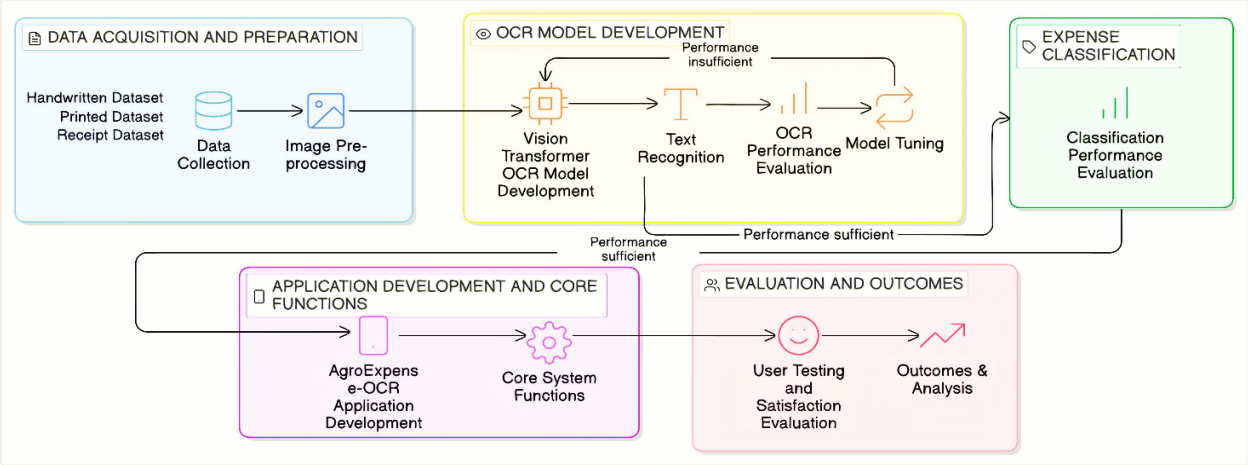

Fig. 1. Research conceptual framework.

During the data collection process, a purposive sampling technique was employed, with an emphasis on farmers who regularly retained agricultural receipts to ensure that the testing process reflected realistic and practical conditions. Additionally, students with foundational knowledge of income and expense accounting were included to support the evaluation of system usability and applicability.

B. Dataset Description

1) Character-level dataset

Two complementary character-level datasets were employed for model pretraining: one comprising handwritten characters and the other containing printed text.

The handwritten dataset includes 41,020 images (224×224 pixels, PNG format) collected from participants aged 20–70 years [9]. Each image represents a single pen-written Thai character, covering 44 consonants, 32 vowels, 4 tonal marks, and 10 numerals (0–9). Samples were contributed by farmers and accountants, capturing diverse handwriting patterns typical of Thai agricultural documentation.

To printed-character dataset was synthetically created using twenty popular Thai and English Google Fonts including Kanit, Prompt, Chakra Petch, Sarabun, Krub, Taviraj, Pridi, Mitr, Bai Jamjuree, Itim, Noto Sans Thai, Noto Serif Thai, Sriracha, Charm, Kodchasan, Mali, Athiti, Chonburi, KoHo, and Thasadith (https://fonts.google.com/?hl=th&lang=en_Latn;th_Thai &script=Thai). The printed-character dataset was synthetically generated using twenty Thai and English Google Fonts in four styles: regular, bold, italic, and underlined. It includes Thai and English alphabets in both uppercase and lowercase, as well as Arabic and Thai numerals (0–9). For non-underlined styles, the dataset also contains Thai vowels, tonal marks, and common receipt symbols such as /, -, %, and ฿. The final dataset comprises 11,740 images in PNG format, encompassing diverse typographic appearances to support robust OCR model. Examples of the character dataset are shown in Fig. 2.

Fig. 2. Example of character dataset.

Together, these datasets were used to pretrain, enabling it to capture fine-grained visual and linguistic representations across Thai and English scripts before full receipt-level training.

2) Receipt-level dataset

The receipt-level dataset contains 3,600 agricultural receipt images collected in Phayao Province, Thailand (2024–2025), including 1,600 printed and 2,000 handwritten samples from farmers and equipment vendors. It consists of rows of product items, including 9,000 printed samples and 6,000 handwritten samples see Table I. Fig. 3 provides examples of printed and handwritten receipts.

TABLE I. DATASET CHARACTERISTICS SUMMARY

Category	Quantity	Description
Total Item Entries	15,000+	Overall number of item lists samples used in the study
Printed Item Entries	9,000	Item lists extracted from printed agricultural receipts
Handwritten Item Entries	6,000	Item lists extracted from handwritten agricultural receipts
Handwritten Characters	41,020	44 Thai consonants, 32 vowels, 4 tonal marks, 10 Thai numerals
Printed Characters	11,740	Thai characters, English characters, numerals, and special symbols (/, −, %, ฿)

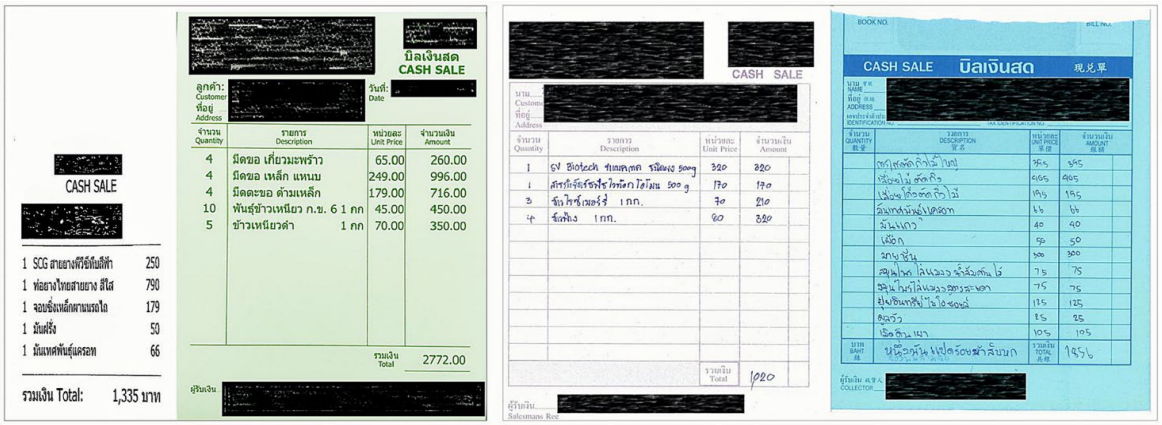

Fig. 3. Example of (a) Printed receipt and (b) Handwritten receipt.

Each receipt features mixed Thai English text such as product names, quantities, and prices. Images were captured under varied lighting, paper textures, and handwriting styles to reflect real-world diversity. This

dataset is used to fine-tune and evaluate the AgroExpense-OCR model for performance across different layouts and recording styles in Thailand's agricultural sector.

Prior to processing by ViT, receipt images undergo an initial preprocessing phase that encompasses noise removal, conversion to grayscale, and resizing. Subsequently, the images are segmented into fixed-size patches, which are then linearly projected into patch embeddings. Each embedding is augmented with positional encodings to preserve spatial order. Within the ViT, the multi-head self-attention mechanism computes contextual relationships among all patches simultaneously, enabling the model to effectively recognize both printed and handwritten text from semi-structured receipts. The final outputs are decoded into characters and stored in the database for subsequent expense categorization and analysis.

3) Ground-truth annotation and quality control

Receipt images were collected from consenting volunteers and scanned at 300–600 dpi. Annotators labeled text regions/lines and key fields in [annotation tool] under written guidelines; a research assistant selected a subset for independent review with senior adjudication. Automated checks verified box-to-transcript consistency, character set validity, and schema completeness; all personally identifiable and location-revealing content was redacted before analysis. The dataset was split into training 70%, validation 15%, and testing 15%, with no vendor overlapping.

4) Data ethics and privacy

Personal identifiers such as farmer names, phone numbers, bank account numbers, vendor names, and addresses were redacted or masked before any processing. All redactions were verified by human reviewers, and only de-identified data were used in subsequent workflows. Privacy safeguards were applied at every stage of data handling.

C. Hardware

Model training and experiments were executed on a workstation with an AMD Ryzen 9 7945HX (2.50 GHz, Radeon Graphics), 64 GB RAM, and an NVIDIA GeForce RTX 4090. Mobile testing was performed on a diverse set of user devices, with application builds supporting both iOS and Android platforms.

IV. Methods

A. Research Design and Framework

This study employed a mixed methods design that integrated quantitative and qualitative techniques. Development proceeded iteratively, with continuous input from agricultural stakeholders guiding each cycle. The framework comprised five consecutive stages: requirements analysis, model construction, application implementation, system integration, and comprehensive evaluation.

The study spanned six months (November 2024–April 2025) in Phayao Province, Thailand, and neighboring areas with intensive field-crop activity. The workflow followed two stages: (i) laboratory development of the OCR model and (ii) field deployment in partnership with farmers. Month 1 centered on a targeted literature review and requirements elicitation alongside initial receipt acquisition; Months 2–3 addressed dataset curation and iterative model training; Month 4 focused on application engineering and end-to-end integration for mobile and web; Month 5 comprised system testing, debugging, and a pilot rollout; and Month 6 finalized user evaluation (usability and recognition accuracy) and overall analysis. This cadence ensured a disciplined transition from evidence-based design to validated field performance within the six-month window.

B. Preprocessing and Layout Analysis

1) Data preprocessing

Prior to processing by the Vision Transformer (ViT), receipt images undergo an initial preprocessing phase that encompasses noise removal, conversion to grayscale, and resizing. Subsequently, the images are segmented into fixed-size patches, which are then linearly projected into patch embeddings. Each embedding is augmented with positional encodings to preserve spatial order. Within the ViT, the multi-head self-attention mechanism computes contextual relationships among all patches simultaneously, enabling the model to effectively recognize both printed and handwritten text from semi-structured receipts. The final outputs are decoded into characters and stored in the database for subsequent expense categorization and analysis.

2) Character-level data preprocessing

The pipeline loads handwriting images from labeled subdirectories, where each folder name encodes the ground-truth text. Images are converted to RGB, resized to 224×224, tensorized, and normalized with ImageNet statistics to provide consistent inputs to the models.

For labels, character-level vocabulary is derived from all annotations and augmented with [BLANK] and [UNKNOW] for CTC decoding. Each label is mapped to an index sequence with its length recorded for batching; empty labels are replaced with [UNKNOW] to ensure valid training samples. This yields a uniform, CTC-compatible representation for both images and text.

3) Receipt-level data preprocessing

The preprocessing procedure begins by normalizing receipt images and separating their primary layout regions. Text-block segmentation is then performed to extract item descriptions, quantities, unit prices, and total amounts. Individual line items are identified using line order information in combination with column boundaries or column-separation lines typically found in receipt structures. This method ensures clear localization and consistent organization of each semantic component, supporting accurate text recognition and subsequent analysis. An example of the extracted text blocks is illustrated in Fig. 4.

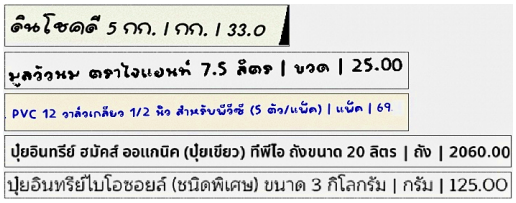

Fig. 4. Example of extracted text blocks.

4) Data augmentation

Researchers apply data augmentation to the main transaction fields, including items, unit price, quantity, and amount, to improve OCR robustness. The augmentation simulates real capture conditions using wrinkles, stains, lighting variation, image blur, resizing, and JPEG compression. For numeric fields, we enforce a consistency rule so that the total amount equals quantity multiplied by unit price. These techniques help the model perform more reliably in real mobile environments.

The dataset for AgroExpense-OCR development agricultural samples, featuring both printed and handwritten forms. These were collected from real farmers' receipts and supplemented with handwritten samples created by students, farmer families, and stakeholders. The dataset includes diverse receipt layouts with Thai text, numbers, and common symbols, making it suitable for reliable OCR training and evaluation.

During pre-processing, several steps were taken to improve data quality. These included noise filtering, resizing and sharpening images, converting them to grayscale for consistency, and segmenting individual characters to prepare data for generating patch embeddings within ViT architecture. These processes effectively reduced recognition errors and improved the accuracy of subsequent OCR model training. These processes effectively reduced recognition errors and improved the accuracy of subsequent OCR model training.

The diversity in the dataset, in terms of both structure and writing styles, highlights its suitability for training Vision Transformer-based OCR models, which need exposure to complex and varied document formats. Including both printed and handwritten receipts enhances representativeness and robustness, especially for Thai farmers who record transactions in both formats. International research stresses that dataset diversity is essential for improving model generalization and reducing recognition errors in OCR applications [16, 21].

C. OCR Model Development on Agricultural Receipts

The training process is conducted in two stages.

Stage 1 involves pretraining the model on a character-level dataset to learn fundamental visual and textual representations.

Stage 2 then fine-tunes the pretrained model on the agricultural receipt dataset to adapt it to domain-specific layouts and content.

The AgroExpense-OCR model adopts Vision Transformer (ViT) architecture because of its ability to convert input images into sequences of patch embeddings. Each patch is linearly projected into a vector, which is then processed through the multi-head self-attention mechanism to capture spatial dependencies across the entire image. This design allows the model to better identify contextual relationships among distant characters compared to traditional Convolutional Neural Networks (CNNs). Hyperparameters such as patch size, embedding dimensions, and the number of attention heads were carefully adjusted to work with both printed and handwritten text. This approach supports the findings of Dosovitskiy *et al.* [21], who showed that ViTs outperform CNNs in various computer vision tasks.

This research presents a ViT-based OCR system for agricultural receipts and evaluates three lightweight models, EfficientViT, FastViT, and MobileViT, to select the most appropriate one for mobile use. The evaluation considers both accuracy and computational efficiency under limited hardware conditions. The study highlights key architectural components such as attention design, token mixing, and model efficiency that affect OCR performance under strict latency and memory requirements. The principles of each model are explained in the following subsections.

1) EfficientViT

EfficientViT is designed as a memory-efficient Vision Transformer that integrates a sandwich-style structure with Cascaded Group Attention (CGA) to minimize attention redundancy and reduce memory overhead while maintaining global feature representation. The model limits the number of attention operations by inserting a single attention layer between Feed-Forward Network (FFN) layers and employs CGA to enable progressive feature interaction across attention heads. This architectural design leads to improved computational efficiency and faster inference compared with standard ViT models. The full mathematical formulation is based on the original EfficientViT framework proposed by Liu *et al.* (https://doi.org/10.1109/CVPR52729.2023.01386) and is directly adopted in this work.

2) FastViT

FastViT adopts a convolution-based RepMixer block to perform token mixing without relying on self-attention, which helps lower computational overhead. For an input feature tensor $X \in \mathbb{R}^{H\times W\times C}$, the original mixing process is defined as:

$$Y = BN(\sigma(DWConv(X))) + X \tag{1}$$

where $DWConv$ is a depth-wise convolution, σ denotes a non-linear activation, and BN is batch normalization. The RepMixer simplifies this operation by removing the activation function and rearranging the computation as

$$Y = DWConv(BN(X) + X) \tag{2}$$

This design enables structural reparameterization, allowing the block to be converted into a single depth-wise convolution at inference time, expressed as $Y = \mathrm{DWConv}(X)$. In this formulation, H and W represent the spatial dimensions and C denotes the number of channels. This simplified inference structure reduces memory access and latency, making FastViT suitable for real-time OCR on mobile platforms [22].

3) MobileViT

MobileViT combines local spatial feature extraction from convolution with global context modeling from transformers. Given a local feature map X_L, the model reshapes it into non-overlapping patches $X_U \in \mathbb{R}^{P \times N \times d}$, where $P = wh$ is the total number of patches, h and w point to the patch height and width, and $N = \frac{HW}{P}$ is the number of pixels in each patch, with H and W representing the feature map dimensions. Each patch is then processed using a transformer to encode global context, formulated as

$$X_G(p) = Transformer\big(X_U(p)\big),\ 1 \le p \le P \quad (3)$$

where $X_G(p)$ servers as the global representation of the p-th patch. After that, the transformed features are folded back to the original spatial resolution and fused with the local features using convolution and feature concatenation [23]. This design allows MobileViT to preserve both local character-level details and global document-level context, resulting in an effective receptive field covering the entire image, which is well suited for OCR tasks on mobile devices.

From a theoretical perspective, EfficientViT, FastViT, and MobileViT represent three distinct strategies for mitigating the quadratic complexity of classical ViTs. EfficientViT achieves complexity reduction through kernelized linear attention and structured group cascading, preserving global dependency modeling with linear scaling [24].

FastViT eliminates self-attention entirely and substitutes it with convolutional token mixing and structural reparameterization, yielding minimal latency and maximal deployment efficiency [22].

MobileViT adopts a hybrid convolution-transformer formulation that balances local stability and global semantic modeling [23].

Consequently, these three models define complementary points in the theoretical design space of efficient transformer-based vision systems for mobile OCR deployment.

D. Field Extraction and Normalization

In the post-OCR stage, extracted text lines are mapped to predefined fields such as date, vendor, item name, quantity, unit price, and total amount using a combination of heuristic rules, probabilistic matching, and regular expressions with locale-aware number and currency parsing. A graph-based alignment approach is then used to merge multi-line items and to resolve column alignment in semi-structured table formats.

E. Expense Data Categorization and Analytics

Automatic categorization of agricultural expenses is essential for the AgroExpense-OCR system to convert raw receipt data into useful business insights. After character recognition by the OCR module, the extracted information is processed using classification algorithms to group expenses into categories such as labor, seeds, chemicals, fertilizers, and machinery costs. This categorization combines rule-based methods and deep learning techniques, especially models, which are effective in capturing contextual relationships between words and numbers in semi-structured receipts.

Once categorized, the data undergoes statistical analysis to summarize expenses at both the category and overall levels, including average costs per category, expense-to-investment ratios, and expenditure trends across production cycles. These analyses offer valuable insights into strategic decision-making by helping farmers visualize spending patterns, identify redundancies, and optimize investment planning. International studies further show that integrating OCR with statistical analytics improves financial management in agriculture and small businesses [14].

F. Evaluation Metrics

The performance of the developed OCR model is assessed using loss, Character Error Rate (CER), and standard classification metrics including Accuracy, Precision, Recall, and F1−Score. The training and validation loss are analyzed to track learning progress and model stability. CER serves as the main metric for character-level recognition by quantifying the differences between predicted outputs and ground-truth text. Accuracy measured the overall percentage of correctly classified characters, while Precision evaluated the correctness of predictions for each identified class. Recall measured the model's ability to retrieve all relevant instances in the dataset, and the F1−Score, as the harmonic means of Precision and Recall, offered a balanced view of accuracy and completeness. Recent studies, such as TrOCR by Li *et al.* (2023) [16], have shown that OCR models based on Vision Transformer architectures outperform traditional CNN-RNN approaches, especially in recognizing text from complex document structures and handwritten inputs. The Attention mechanism within ViTs allows the model to effectively capture long-range dependencies, leading to higher Accuracy and F1-scores in real-world agricultural receipt recognition scenarios.

Accuracy, Precision, Recall, and F1−Score are used to assess field-level predictions for structured information such as item names, quantities, unit prices, and total amounts, providing an overall view of recognition reliability in practical use.

G. AgroExpense-OCR Application Development

1) Mobile application development

AgroExpense-OCR was developed as a mobile application using Flutter and Android Studio, providing cross-platform compatibility for both Android and iOS devices. This design choice enhances accessibility for farmers across diverse groups. The application emphasizes a user-friendly interface, aligning with studies highlighting the role of mobile technologies in advancing digital farm management and reducing the burden of manual data entry.

2) Web application development

To support system-level data storage and analysis, AgroExpense-OCR was built using Node.js, Express, and MySQL, ensuring efficient handling of large-scale

databases. This stack enables real-time processing and smooth integration with the OCR module. International research confirms that integrating database-driven web applications with OCR systems improves cost analysis accuracy and enhances decision-making in agriculture.

H. Testing and Evaluation of System Usability

1) OCR model performance testing

The performance of the OCR model built with Vision Transformer (ViT) was evaluated on a test dataset separate from the training and validation sets. Standard metrics, including Accuracy, Precision, Recall, and F1-score, were used to measure both the correctness and completeness of character recognition. The ViT was selected for its ability to more effectively handle semi-structured documents than traditional CNN-RNN approaches. Recent research has shown that TrOCR, a ViT-based OCR, achieves better results in both printed and handwritten text recognition compared to conventional methods [16].

2) Application usability testing with real users

To assess the operational efficacy of AgroExpense-OCR, usability testing was performed with actual users under conditions that closely mimic real-world scenarios. In accordance with the Usability Testing Framework, the evaluation encompassed efficiency, effectiveness, and user satisfaction. Testing involving farmers and stakeholders confirmed that the system successfully minimized manual data entry and enhanced access to cost analysis. Research on mobile agricultural applications has verified that a usability-centered design markedly increases user acceptance and adoption within agricultural environments [25].

3) User satisfaction assessment

An assessment of user satisfaction was conducted to gather the perspectives of farmers and stakeholders regarding system utilization. Structured questionnaires and interviews were employed, concentrating on recognition accuracy, usability, and perceived advantages in managing agricultural expenses. The findings were utilized to enhance the system, ensuring better alignment with user requirements. Research within the field of Human-Computer Interaction (HCI) has consistently shown that evaluations of user experience and satisfaction are crucial determinants of the success of digital agricultural systems [26].

V. Research Result

This section presents the results from designing, developing, and evaluating the AgroExpense-OCR system. The findings are organized to align with the research goals and methodology, offering a comprehensive overview of the outcomes at different study stages. First, the characteristics and size of the agricultural dataset, including printed and handwritten samples, are described, followed by the results of data pre-processing and preparation. The following section presents OCR performance metrics along with example results from real data.

Additionally, the results of the AgroExpense-OCR application development are presented, covering both mobile (Android/iOS) and web platforms, with a focus on core features such as OCR recognition, automatic categorization, budgeting, and notifications. This section also includes findings related to expense categorization and analytics, highlighting the effectiveness of automatic classification and the creation of statistical summaries and visualizations.

Finally, user evaluation results are discussed, incorporating feedback from farmers and stakeholders to assess usability, effectiveness, and areas for improvement. These outcomes demonstrate both the technical capabilities of the system and its practical value in improving agricultural cost management through automated digitization and analytics.

A. Parameters Setting

This subsection shows the main parameter settings used for training EfficientViT, FastViT, and MobileViT for OCR comparison. The detailed configurations are represented in Table II.

TABLE II. Parameter Settings for OCR Training Models

Parameter	EfficientViT	FastViT	MobileViT
Backbone	*efficientvit_b1*	*fastvit_sa12*	*mobilevit_s*
Input Size	224 × 224	224 × 224	224 × 224
Batch Size	32	32	64
Learning Rate	1×10^{-5}	1×10^{-5}	1×10^{-5}
Optimizer	AdamW	AdamW	AdamW
Weight Decay	–	–	5×10^{-2}
Epochs	100	100	100
Loss Function	CTC Loss	CTC Loss	CTC Loss
Decoding	Greedy and Beam	Greedy and Beam	Greedy and Beam

Table II presents the core training settings of EfficientViT, FastViT, and MobileViT used in this work. The three models are trained using AdamW and CTC loss with the same learning rate and number of training epochs to ensure a fair evaluation. Input resolution and batch size are chosen to balance computational cost and recognition performance. The backbone architecture follows the official implementations provided in the original model repositories. Greedy decoding is applied for real-time mobile inference, whereas beam search is reserved for evaluation purposes.

The exported checkpoint sizes are consistent with the model architecture. FastViT, which is the best-performing model in this study, has the largest size at about 40 MB, while EfficientViT and MobileViT remain lightweight at around 18 MB and 20 MB. Despite its larger size, FastViT still falls within an acceptable range for mobile OCR deployment.

B. Text Recognition Performance of the Vision Transformer-Based OCR Model

In text recognition performance process of the ViT-based OCR models, including EfficientViT, FastViT, and MobileViT, is evaluated using training and validation loss

to illustrate learning trends, stability, and convergence behavior, as shown in Fig. 5.

1) Training loss and validate loss

The findings reveal distinct training characteristics across FastViT, EfficientViT, and MobileViT. All models show a rapid decline in training loss during the first 5–10 epochs, dropping from roughly 1.0–2.5 to below 0.2 and later stabilizing around 0.02–0.05. EfficientViT and MobileViT maintain consistently stable validation loss values, typically under 0.1 for both printed and handwritten inputs. In contrast, FastViT displays notable instability when learning handwritten samples, with several validation peaks rising above 2.5 between epochs 20–50. Nevertheless, FastViT converges the quickest and ultimately achieves the lowest CER among the models.

For mobile OCR, FastViT is preferred for its low latency after Structural Reparameterization. MobileViT is slower but stable, while EfficientViT suits CPU devices. In practice, use Greedy decoding for printed text and Beam Search with a lightweight language model for handwritten input to balance speed and accuracy.

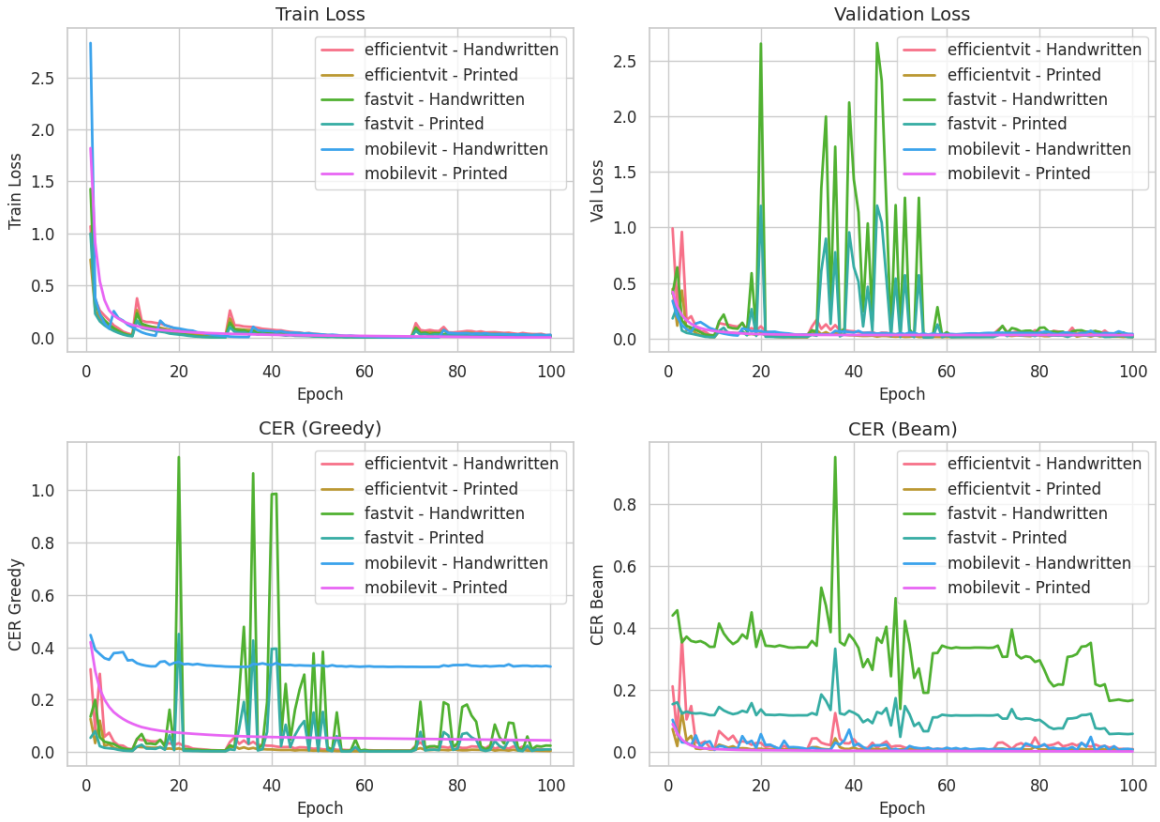

Fig. 5. Comparison of training loss, validation loss, and CER of EfficientViT, FastViT, and MobileViT during OCR model training.

2) CER greedy and CER beam

The numerical results clearly highlight the differences in convergence speed and recognition accuracy among the three architectures. FastViT demonstrates the fastest learning behavior, as its CER on printed text falls below 5% within the first 20–30 epochs. Nevertheless, the training process shows temporary instability, where the CER rises to around 20–30% at certain stages before recovering. In comparison, MobileViT and EfficientViT require a longer training period, typically around 50–70 epochs, to reach comparable accuracy. After full convergence, all models achieve low CER on printed data, normally within 3–5%. However, for handwritten text, the error rate remains considerably higher, approximately 15–35%, indicating that handwritten recognition is still much more difficult than printed text.

For decoding methods, Beam Search clearly gives better results than Greedy Search for handwritten text. In the Greedy graph, FastViT has many sharp spikes. In the Beam graph, these spikes are reduced. The overall handwritten CER is also lower. For printed text, the difference between Greedy and Beam is very small. Both methods give near-zero CER after convergence. This means Greedy decoding is already good enough for printed documents. Beam Search is mainly important for handwritten recognition.

3) Character-level test performance

This section reports the Character-Level Test Performance of the models, evaluated using test loss and CER with greedy and beam search decoding on 20,467 item-derived samples, represent in Table III.

TABLE III. CHARACTER-LEVEL TEST PERFORMANCE

Model	Test Loss	CER Greedy	CER Beam
EfficientViT	0.0404	0.0089	0.0093
FastViT	0.0314	0.0086	0.0091
MobileViT	0.0925	0.1647	0.0917

The experimental results shown in Table III, FastViT deliver the strongest performance, achieving the lowest test loss (0.03144) and CER values of 0.0086 with greedy decoding and 0.0091 with beam search. EfficientViT produces slightly higher results, with a test loss of 0.0404 and CER of 0.0089 (greedy) and 0.0093 (beam), while maintaining stable performance. In comparison, MobileViT exhibits considerably higher error rates, with a test loss of 0.09258 and CER of 0.1647 under greedy decoding, which decreases to 0.0917 when using beam search.

Additional error analysis indicates that many recognition errors arise from visually similar symbols, especially the English lowercase letter “b”, the Arabic numeral “6”, and the Thai vowel “ใ”. Other confusing pairs are also found, like “l (L) –I–1 (One)”, “0–6”, “s–5”, and Thai characters like “เ–แ” and “ว–ล”. In handwritten data, these characters often have very similar stroke patterns. This increases visual confusion and leads to misclassification.

OCR errors observed in both handwritten and printed receipts mainly arise from diverse handwriting styles, inadequate image quality, irregular character spacing, and background noise or visual artifacts. These issues are especially pronounced in mobile OCR applications, where variations in writing and image capture conditions are difficult to control, leading to challenges in reliable text recognition.

For mobile deployment, FastViT is the best choice for real-time OCR because it gives very low CER for printed text and has fast inference. For receipts and printed documents, FastViT with Greedy Search is recommended for lowest latency. For handwritten notes, FastViT should be used with Beam Search to reduce errors. EfficientViT is suitable for low-power devices, and MobileViT is stable but slower. This setup gives a good balance between speed and accuracy for real mobile OCR applications.

C. Test Performance of Expense Classification

Items dataset is categorized into two distinct groups: (i) Management and Operations, which include machinery, labor, and transportation; and (ii) Manufacturing Costs, covering fertilizers, pesticides, seeds, and miscellaneous supplies. Real-world estimations were carried out on a dataset of 2,250 product items obtained from 375 receipts, with the items evenly categorized into two classes of 1,125 items each.

FastViT is used as the backbone for mobile product classification. The model is evaluated using Accuracy, Precision, Recall, and F1−Score. The results indicate that FastViT provides high classification accuracy and stable class-wise performance, confirming its suitability for product classification, shown in Table IV.

TABLE IV. OCR MODEL EVALUATION METRICS

Metric	Printed Receipts	Handwritten Receipts
Accuracy	96.80%	92.40%
Precision	95.70%	91.20%
Recall	96.30%	90.80%
F1-score	96.00%	91.00%

Table IV Evaluation Metrics of the Vision Transformer-Based OCR Model. This table summarizes the evaluation results of the OCR model built on the Vision Transformer architecture, comparing performance between printed and handwritten receipts. The results indicate that printed receipts achieved high scores across all metrics, with Accuracy, Precision, Recall, and F1−Score exceeding 95%, reflecting consistent and stable recognition performance. Handwritten receipts, while slightly lower (around 91–92%), remained within a practically acceptable range. The performance gap can be attributed to handwriting variability and irregular text alignment. Nevertheless, the results confirm that the model effectively handles diverse and semi-structured receipt formats, outperforming traditional CNN-RNN based methods.

In practical deployment, the OCR engine occasionally introduces minor inaccuracies in Thai diacritics, including tone marks and the *karan*, with the fourth tone mark (◌๋) often misinterpreted as the Thai numeral seven (๗) or placed incorrectly. These issues produce surface-level variants such as น้ำ *(water)* appearing as น้า or น ◌ำ◌้, and ปุ๋ย *(fertilizer)* appearing as ปุย or ปุ๋ย. Despite these orthographic deviations, the effect on downstream product classification is minimal because the core lexical structure remains recognizable and continues to convey sufficient semantic information for reliable categorization in real-world retail OCR pipelines.

D. Application Development Outcomes

The AgroExpense-OCR application was developed in two versions: a mobile application and a web application. The mobile application, constructed utilizing Flutter and Android Studio, was designed to operate on both Android and iOS platforms. Its interface prioritized simplicity and user-friendliness for farmers, incorporating essential functionalities such as receipt scanning, user registration, and visualization of expense reports through charts and tables. Testing with sample users demonstrated that the mobile application facilitated smooth execution of key functions and effectively addressed the practical needs associated with recording agricultural expenses.

The web application developed using Node.js, Express, and MySQL, served as the central repository and processing hub. It supported data storage originating from the mobile application, enabled automated expense categorization, and facilitated the generation of analytical reports. Integration between the web and mobile platforms ensured that users could access data seamlessly via online and portable devices.

Regarding core functionalities, the developed system implemented the following features: (1) user registration and authentication; (2) receipt management and data extraction using an Optical Character Recognition (OCR) service; (3) category management for organizing expense data; (4) budget management for monitoring expenditures; and (5) data visualization for financial analysis. These outcomes underscore the efficacy of AgroExpense-OCR in promoting systematic and sustainable expense management for farmers.

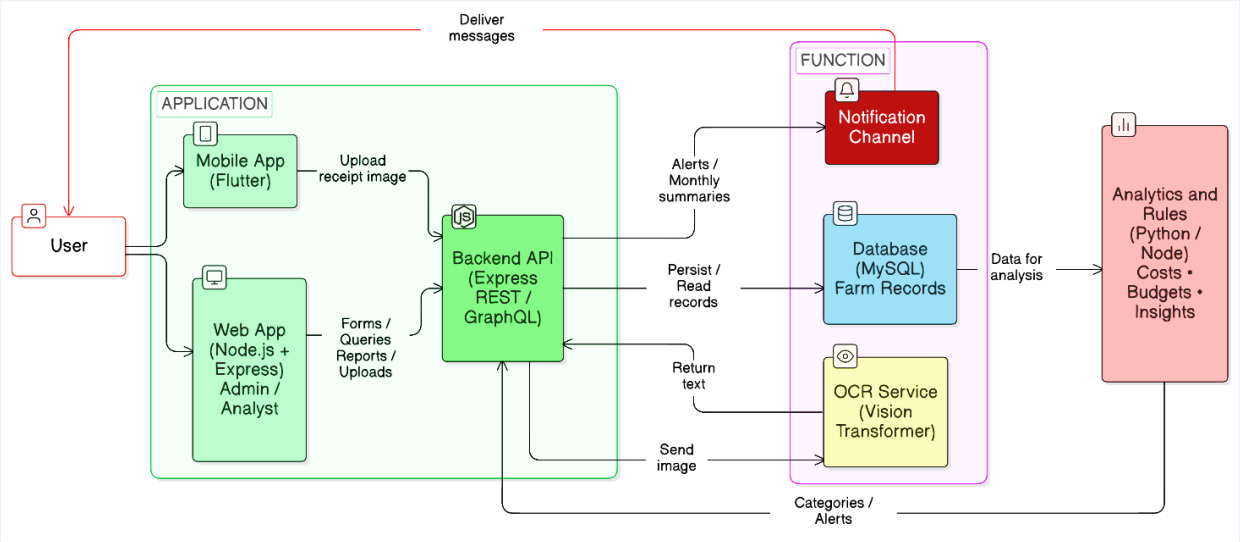

Fig. 6. AgroExpense-OCR — System architecture and connections.

Fig. 6 shows the system architecture of AgroExpense-OCR, which consists of four main layers: Users, Mobile/Web Clients, Backend API, and Core Services (OCR, Database, Analytics). Farmers and stakeholders interact with the system through the mobile app (built with Flutter for Android and iOS) or the web app (developed using Node.js and Express). Data inputs, such as receipt images, are sent to the Backend API.

The Backend API processes the data and communicates with the OCR Engine, built using a Vision Transformer, to convert textual information into structured digital data. The extracted data are stored in the Database and sent to the Analytics & Rules module for expense categorization, budgeting, and notifications. This workflow enables users to access past data, analyze farming expenses, and make informed financial decisions effectively.

Table V shows the primary responsibilities and interfaces of each core component throughout the entire pipeline. Users upload receipt images and receive results or alerts through the Mobile App or Web App. Both clients communicate with the Backend API Express, which manages REST/GraphQL, authorization, routing, service orchestration, and calls the OCR Service Vision Transformer to convert receipt images into structured Text JSON. Data is stored in the Database (MySQL) and used by the Analytics & Rules module for categorization, budgeting, and notifications, which are sent through the Notification Channel FCM/Email/SMS. The table explains data flow and system boundaries, supporting consistent testing, monitoring, and future scaling. An example of application is shown in Fig. 7.

TABLE V. SYSTEM COMPONENTS AND RESPONSIBILITIES

Component	Technology/Stack	Primary Responsibilities	Data In/Out	Key Interfaces	Notes
Users	–	Initiate actions; capture receipts; review analytics	Uploads receipt images; receives results/alerts	Mobile App, Web App	Farmers & stakeholders
Mobile App	Flutter (Android/iOS)	Capture images, local validation, send to backend, show results	Image files, metadata; displays OCR & analytics	Backend API	Camera access; offline cache (optional)
Web App	Node.js + Express (UI)	Manage accounts, review/edit records, dashboards	Forms/queries; shows aggregated analytics	Backend API	Administrative functions
Backend API	Express (REST/GraphQL)	Auth, routing, orchestration; call OCR; persist/retrieve data	JSON requests/responses	Mobile/Web, OCR Service, Database, Analytics	Rate limiting, logging
OCR Service	Vision Transformer	Text extraction from receipts; post-processing	Receipt image → Text JSON	Backend API	Handles printed & handwritten Thai text
Database	MySQL	Store users, receipts, items, categories, budgets, logs	CRUD operations; aggregates	Backend API, Analytics	Backup/replication policy
Analytics & Rules	Python/Node jobs	Auto-categorization, budgeting rules, notifications	Reads records; outputs categories, budgets, alerts	Backend API, Database	Batch/near-real-time
Notification Channel	FCM/Email/SMS	Delivery alerts & summaries	Messages/alerts	Backend API	Budget breach, anomalies

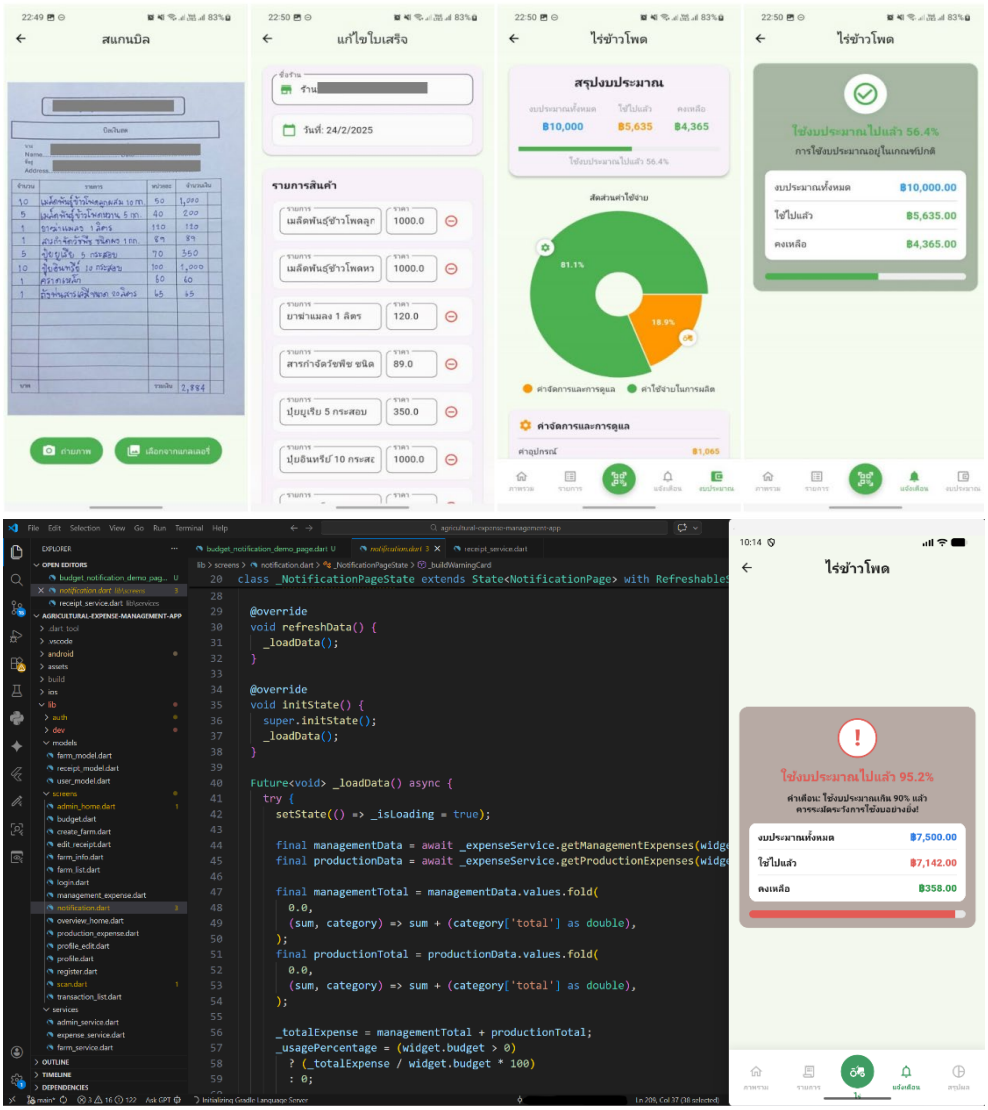

Fig. 7. Example of application.

E. *Results of Expense Categorization and Analytics*

The study results indicate that AgroExpense-OCR effectively enabled automatic categorization of agricultural expenses using receipt data processed by the OCR module and subsequent classification algorithms. The system achieved high accuracy in grouping expenses into categories such as labor, agricultural inputs (seeds, fertilizers, and chemicals), fuel, and machinery. This automatic classification significantly reduced manual data entry workload and organized the data for efficient statistical analysis.

Regarding statistical analysis, the system provided summaries of expenses at both the category and overall levels, including average costs per category, expense-to-investment ratios, and trends across production cycles. The results were presented through tables, charts, and interactive dashboards, making it easy for users to visualize spending patterns. These findings are consistent with international research, which confirms that integrating OCR with data analytics offers systematic support for strategic decision-making in agriculture.

F. *User Evaluation and Satisfaction Assessment*

The user evaluation of AgroExpense-OCR was conducted among farmers and agricultural stakeholders. The findings showed that most participants were highly satisfied with the system, especially with the accuracy of character recognition and the convenience of automated expense categorization. These features helped reduce the manual record-keeping burden and sped up cost analysis.

Regarding usability, participants confirmed that the system interface was intuitive and easy to navigate, even for those with limited technical skills. However, several improvement suggestions were made, including adding multilingual support, offering more customizable dashboards, and enabling interoperability with external accounting systems to enhance financial management.

Overall, the evaluation demonstrated that AgroExpense-OCR not only met farmers' practical needs for recording and analyzing expenses but also increased confidence in cost management, supporting efficiency and sustainability in agricultural production.

Table VI presents the quantitative assessment of user satisfaction across different areas. Results indicate that OCR Recognition Accuracy scored one of the highest at 4.6/5, while Automatic Categorization was rated as the most valuable feature with 4.7/5, showing strong user confidence in the system's main functions. Usability of the Mobile Application (4.5) and Web Application (4.4) were both rated positively, implying ease of use overall, though participants suggested further improvements like more flexible dashboards and multilingual options. Budgeting Features (4.3) and Notifications (4.2) were seen as useful but identified as areas for potential enhancement. Overall, users rated Overall Satisfaction at 4.5/5, indicating high acceptance and confidence in the system's effectiveness.

TABLE VI. USER EVALUATION AND SATISFACTION ASSESSMENT

Evaluation Dimension	Mean Score (out of 5)	Interpretation
OCR Recognition Accuracy	4.6	High accuracy in both printed and handwritten receipts
Usability of Mobile Application	4.5	Easy-to-use interface, suitable for farmers
Usability of Web Application	4.4	Web dashboards intuitive but with suggestions for customization
Usefulness of Automatic Categorization	4.7	Categorization features considered highly useful
Effectiveness of Budgeting Features	4.3	Budgeting tools effective but need expansion
Notification and Alerts	4.2	Alerts functional though improvement suggested
Overall Satisfaction	4.5	Users generally very satisfied with system performance

VI. DISCUSSION

The research results show that using Vision Transformers (ViTs) for agricultural receipt OCR greatly enhances data recognition efficiency. FastViT is identified as the most robust architecture for mobile and web applications, offering superior convergence and the lowest overall CER. While all models excel with printed text, the inherent complexity of handwritten recognition is effectively managed through optimized decoding strategies that stabilize performance. Consequently, FastViT is deployed as the primary model to optimally balance latency and precision. This system leverages a context-aware decoding approach to ensure high efficiency for printed documents while maintaining reliable accuracy for handwritten inputs. Furthermore, the proposed methodology streamlines the recognition pipeline by utilizing an integrated character set, allowing consonants, vowels, and numerals to be processed as a single entity across both printed and handwritten formats. This approach presents a distinct advantage over the framework established by Nuankaew *et al.* [9], which is focused exclusively on handwritten text and relies on the pre-classification of data into separate components. By eliminating the requirement for data partitioning and extending the operational scope to include printed media, the current model provides a more versatile and efficient solution for mobile OCR.

The test performance of expense classification achieved over 96% accuracy for printed receipts and around 91–92% for handwritten receipts, exceeding traditional CNN-RNN methods for semi-structured agricultural documents. Additionally, the AgroExpense-OCR app, designed for both mobile and web platforms, helps farmers digitize, categorize, and analyze expenses systematically, reducing manual bookkeeping while improving cost planning and financial management in agriculture. However, recognition accuracy drops when processing low-quality receipts or highly variable handwriting, which aligns with Nimsuk *et al.* [12], who pointed out the difficulties of Thai handwriting recognition. These findings indicate strong potential for integrating OCR technology into smart farm accounting and digital agricultural management systems to support farmers' long-term financial stability.

Researchers found that recent international studies on precision agriculture closely match these findings. A comprehensive review covered the use of ViTs in tasks like classification, detection, and segmentation, highlighting their strengths and limitations, especially regarding data needs and model interpretability [27, 28]. This review affirms the potential of ViTs in complex agricultural settings, supporting your research focus.

An interesting study proposed a hybrid knowledge transfer framework from Swin Transformer to MobileNetV3, allowing efficient deployment on resource-limited IoT devices. Despite significant reductions in computational costs (GFLOPs) and inference time, the model kept high accuracy (92.4% for MobileNetV3 vs. 95.9% for Swin-L), making it a promising choice for real-world precision agriculture tasks [29].

In summary, these results place your work at the forefront of global discussions: ViTs are not only effective for OCR in agricultural cost tracking but also align with broader trends in smart agricultural image analysis. There are strong reasons to expand your application toward efficient, real-time IoT-enabled systems.

VII. CONCLUSION

This research concludes that using the Vision Transformer (ViT) for agricultural receipt OCR significantly improves recognition accuracy and reliability for both printed and handwritten text. Moreover, the AgroExpense-OCR application, available on mobile and web platforms, effectively eases farmers' bookkeeping and expense tracking efforts, while enabling more precise financial planning and decision-making.

The text recognition performance of FastViT, EfficientViT, and MobileViT were evaluated on both printed and handwritten datasets. Among the three architectures, FastViT demonstrates the fastest convergence and the lowest final CER, making it the most suitable model for real-time OCR applications. Although FastViT exhibits temporary instability when trained on handwritten data, this issue is effectively mitigated through Beam Search decoding, which enhances recognition stability and reduces error rates. Its structural

reparameterization further enables a favorable balance between inference latency and accuracy, supporting practical mobile OCR deployment. While handwritten text remains more challenging than printed text due to variations in writing style and image quality, the proposed decoding strategy substantially alleviates these difficulties. In addition, expense classification results based on the OCR outputs highlight the value of vision-based architecture for analyzing semi-structured agricultural documents in this study.

In addition to OCR performance, this study validates FastViT as an effective backbone for expense classification, achieving high accuracy and stable performance across management, operations, and manufacturing cost categories. In line with recent international studies, the results confirm the suitability of advanced vision transformer architectures for analyzing semi-structured agricultural documents. The AgroExpense-OCR system was successfully deployed on mobile and web platforms to support efficient receipt capture, expense extraction, and centralized data processing, with evaluations indicating strong user acceptance, reliable recognition performance, and practical value for agricultural expense management.

This study's main contribution is the introduction of a unified OCR framework capable of handling both printed and handwritten text via a single integrated character set, thereby removing the need for pre-classification or segmented processing. Compared to approaches limited to handwritten text, the proposed method offers improved adaptability and efficiency. The integration of an optimized Vision Transformer model with context-aware decoding strategies delivers a scalable and application-ready solution for agricultural receipt digitization in real operational settings.

Suggestions for future work include expanding the dataset to include a wider variety of receipts and handwriting styles, which would improve the model's robustness. Integrating the system with IoT or edge devices could facilitate real-time application in actual farming environments. Additionally, researchers will explore extending the AgroExpense-OCR framework to support multilingual receipt recognition through character set adaptation and retraining, enabling broader applicability across diverse agricultural contexts. This extension would complement its integration with smart farm accounting systems and agricultural financial platforms to further enhance decision-making and sustainability.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Wongpanya Nuankaew conceptualized the study and designed the methodology; developed and validated the research procedures; implemented the model; conducted data analysis and interpretation; and contributed to drafting, reviewing, and editing the manuscript. Chanapa Phikhason collected and curated the data; developed the database and supporting systems; managed data for analysis; and contributed to manuscript preparation. Watthana Kayowaen implemented system components for data processing and cleaning; performed data quality checks and resolved inconsistencies; generated descriptive summaries; and assisted in preparing figures, tables, and manuscript formatting. Tatsaneewan Yenwattana organized and documented the dataset; maintained version control for data and system outputs; supported data integration and preprocessing; and assisted with literature review and manuscript preparation. Pratya Nuankaew contributed to study conceptualization and design; provided methodological guidance in data science and machine learning; supported model development and validation; reviewed and verified data analysis and interpretation; served as the corresponding author, managing all submission and correspondence-related tasks; coordinated communication among authors and with the journal; and contributed to writing, reviewing, and editing the manuscript; All authors reviewed and approved the final version.

REFERENCES

[1] C. J. Lin, Y. C. Liu, and C. L. Lee, "Automatic receipt recognition system based on artificial intelligence technology," *Applied Sciences*, vol. 12, no. 2, 853, Jan. 2022. doi: 10.3390/app12020853

[2] J. Antonio, A. R. Putra, H. Abdurrohman, and M. S. Tsalasa, "A survey on scanned receipts OCR and information extraction," in *Proc. International Conference on Document Analysis and Recognit*, Jerusalem, Israel, 2022, pp. 29–30.

[3] S. A. Francis and M. Sangeetha, "A comparison study on optical character recognition models in mathematical equations and in any language," *Results in Control and Optimization*, vol. 18, 100532, Mar. 2025. doi: 10.1016/j.rico.2025.100532

[4] S. Raza, M. Farooq, U. Farooq, H. Karamti, T. Khurshaid, and I. Ashraf, "A convolutional neural network based optical character recognition for purely handwritten characters and digits," *CMC*, vol. 84, no. 2, pp. 3149–3173, 2025. doi: 10.32604/cmc.2025.063255

[5] N. Anakpluek, W. Pasanta, L. Chantharasukha, P. Chokratansombat, P. Kanjanakaew, and T. Siriborvornratanakul, "Improved tesseract optical character recognition performance on Thai document datasets," *Big Data Research*, vol. 39, 100508, Feb. 2025. doi: 10.1016/j.bdr.2025.100508

[6] 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," *Neurocomputing*, vol. 556, 126702, Nov. 2023. doi: 10.1016/j.neucom.2023.126702

[7] N. Nithisopa and T. Panboonyuen, "DOTA: Deformable optimized transformer architecture for end-to-end text recognition with retrieval-augmented generation," in *Proc. 2025 17th International Conference on Knowledge and Smart Technology (KST)*, 2025, pp. 301−306.

[8] X. Luan, J. Zhang, M. Xu, W. Silamu, and Y. Li, "Lightweight scene text recognition based on transformer," *Sensors (Basel)*, vol. 23, no. 9, 4490, May 2023. doi: 10.3390/s23094490

[9] W. S. Nuankaew, C. Phikhason, W. Kayowaen, T. Yenwattana, P. N. Ngium, and P. Nuankaew, "Harnessing AI for agriculture: Thai handwriting recognition for agricultural accounting and management using vision transformers," in *Proc. Int. Conf. Digit. Arts, Media Technol., DAMT ECTI North. Sect. Conf. Electr., Electron., Comput. Telecommun. Eng., NCON*, Institute of Electrical and Electronics Engineers Inc., 2025, pp. 12–17. doi: 10.1109/ECTIDAMTNCON64748.2025.10962093

[10] M. S. Basir, D. Buckmaster, A. Raturi, and Y. Zhang, "From pen and paper to digital precision: A comprehensive review of on-farm recordkeeping," *Precision Agric*, vol. 25, no. 5, pp. 2643–2682, Oct. 2024. doi: 10.1007/s11119-024-10172-7

[11] R. Chamchong, U. Saisangchan, and P. Pawara, "Thai handwritten recognition on BEST2019 datasets using deep learning," in *Proc. International Conference on Multi-disciplinary Trends in Artificial Intelligence*, 2021, pp. 152–163.

[12] N. Nimsuk, N. Thumpaiboon, and W. Phuangsri, "Offline handwriting recognition of Thai characters using multiple deep neural networks," in *Proc. 2023 3rd International Symposium on Computer Technology and Information Science (ISCTIS)*, July 2023, pp. 780–785. doi: 10.1109/ISCTIS58954.2023.10213205

[13] K. Khunratchasana and T. Treenuntharath, "Thai digit handwriting image classification with convolutional neural networks," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 27, no. 1, pp. 110–117, July 2022. doi: 10.11591/ijeecs.v27.i1.pp110-117

[14] J. M. Yu, H. J. Ma, and J. L. Kong, "Receipt recognition technology driven by multimodal alignment and lightweight sequence modeling," *Electronics*, vol. 14, no. 9, 1717, Jan. 2025. doi: 10.3390/electronics14091717

[15] R. Atienza, "Vision transformer for fast and efficient scene text recognition," in *Proc. International Conference on Document Analysis and Recognition*, 2021, pp. 319–334. doi: 10.1007/978-3-030-86549-8_21

[16] M. Li *et al.*, "TrOCR: Transformer-based optical character recognition with pre-trained models," in *Proc. AAAI Conference on Artificial Intelligence*, vol. 37, no. 11, June 2023, pp. 13094–13102. doi: 10.1609/aaai.v37i11.26538

[17] A. Abdallah *et al.*, "CORU: Comprehensive post-OCR parsing and receipt understanding dataset," arXiv preprint, arXiv:2406.04493, 2024.

[18] S. Dutta, S. Adhikary, and A. D. Dwivedi, "VisFormers—Combining vision and transformers for enhanced complex document classification," *Machine Learning and Knowledge Extraction*, vol. 6, no. 1, pp. 448–463, Mar. 2024. doi: 10.3390/make6010023

[19] Z. Liu, R. Song, K. Li, and Y. Li, "From detection to understanding: A systematic survey of deep learning for scene text processing," *Applied Sciences*, vol. 15, no. 17, 9247, Jan. 2025. doi: 10.3390/app15179247.

[20] S. Geng, Z. Zhu, Z. Wang, Y. Dan, and H. Li, "LW-ViT: The lightweight vision transformer model applied in offline handwritten Chinese character recognition," *Electronics*, vol. 12, no. 7, 1693, Jan. 2023. doi: 10.3390/electronics12071693.

[21] A. Dosovitskiy *et al.*, "An image is worth 16 × 16 words: Transformers for image recognition at scale," arXiv preprint, arXiv:2010.11929, 2020.

[22] P. K. Anasosalu Vasu, J. Gabriel, J. Zhu, O. Tuzel, and A. Ranjan, "FastViT: A fast hybrid vision transformer using structural reparameterization," in *Proc. 2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2023, pp. 5762–5772. doi: 10.1109/ICCV51070.2023.00532

[23] S. Mehta and M. Rastegari, "MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer," arXiv preprint, arXiv:2110.02178, 2021.

[24] X. Liu, H. Peng, N. Zheng, Y. Yang, H. Hu, and Y. Yuan, "EfficientViT: Memory efficient vision transformer with cascaded group attention," in *Proc. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition* 2023, pp. 14420–14430. doi: 10.1109/CVPR52729.2023.01386

[25] A. Tripathi, A. Jain, A. K. Singh, P. Choudhary, K. K. Mishra, and P. C. Vashist, "The internet of things in agriculture for sustainable rural development," *AI, Edge and IoT-Based Smart Agriculture*, 2022, pp. 157−170.

[26] U. Ibrahim and A. Danmaigoro, "Human-computer interaction in agricultural user interfaces," *International Journal of Applied and Scientific Research*, vol. 2, no. 2, pp. 187–198, Feb. 2024. doi: 10.59890/ijasr.v2i2.1381

[27] M. S and G. R, "Plant leaf disease detection using vision transformers for precision agriculture," *Sci. Rep.*, vol. 15, 22361, July 2025. doi: 10.1038/s41598-025-05102-0

[28] S. Mehdipour, S. A. Mirroshandel, and S. A. Tabatabaei, "Vision transformers in precision agriculture: A comprehensive survey," arXiv preprint, arXiv:2504.21706, 2025.

[29] S. Mugisha, R. Kisitu, and F. Tushabe, "Hybrid knowledge transfer through attention and logit distillation for on-device vision systems in agricultural IoT," arXiv preprint, arXiv:2504.16128, 2025.