Leveraging OCR-Driven Information Extraction for Accurate Product Type Classification from Thai Receipt Data: An Ensemble Learning Approach

Wongpanya S. Nuankaew[1[0000-0003-3805-9529]], Apitarat Autarach[2[0009-0004-5392-2321]], Teerapakorn Meesri[3[0009-0009-8706-2565]], and Pratya Nuankaew[4[0000-0002-3297-4198]]

[1, 2, 3, 4] School of Information and Communication Technology, University of Phayao, Phayao, 56000, Thailand
pratya.nu@up.ac.th

Abstract. This study investigates OCR-driven information extraction and ensemble learning for product type classification from Thai receipt data to enhance family expense management. Using 1,305 receipt images from Thailand, we extracted and preprocessed 5,087 product names across five categories. We compared base classification algorithms with ensemble learning algorithms, focusing on their performance in handling OCR-extracted Thai text. Results demonstrated the superiority of ensemble methods, particularly Majority Voting and Extra Trees, in classifying product types. Majority Voting achieved a weighted average F1-score of 91.74% and accuracy of 91.92%, while Extra Trees recorded the highest overall accuracy at 92.05%. This study contributes to the field by addressing the unique challenges of Thai language OCR and product classification, offering insights into effective ensemble learning strategies for receipt data analysis.

Keywords: Ensemble Learning, Family Expense Management, OCR-Driven Information Extraction, Product Classification, Receipt Data.

1 Introduction

In the era of digital personal finance, receipt data has become a powerful tool for families seeking to comprehend and manage their household expenses. By analyzing detailed information from receipts, families can gain valuable insights into their spending patterns, feelings about purchasing products [1], budget allocation, and consumption trends. The increasing availability of digitized receipt data through Optical Character Recognition (OCR) technology has opened new possibilities for more comprehensive and accurate expense tracking, particularly in multilingual contexts possibilities [2] like Thailand.

Receipt data is crucial for family expense management as it provides granular information about purchased products, including categories, quantities, and prices. This detailed view allows families to track their spending across various classifications, show the identity areas for potential savings, and make informed decisions about their financial habits. Moreover, accurate product classification from receipts can help automate the categorization of expenses, conserve time, and reduce errors in personal budgeting. However, extracting and classifying product information from OCR-processed receipt data [3, 4] presents several challenges, including handling inconsistent formatting, dealing with OCR errors, and accurately categorizing diverse product descriptions. This research proposes a model using ensemble learning techniques for product classification from receipt data. Ensemble learning, which combines multiple machine learning algorithms, shows promise in improving classification accuracy by leveraging the strengths of different approaches.

This research addresses several key challenges: the complexity of OCR for the Thai language, characterized by its unique character set and the absence of word boundaries; the diversity in product naming conventions found on Thai receipts; and the necessity for accurate and efficient classification of products into relevant expense categories. By tackling these challenges, this study aims to advance automated expense management, particularly for Thai-language receipts, while contributing to the broader fields of multilingual OCR and text classification. Specifically, the research introduces base classification models and employs ensemble learning techniques to accurately classify product types from receipt data, aiding family expense management. The model is trained on a dataset meticulously curated from volunteers and receipt data donors in Phayao Province, Thailand, ensuring a diverse and representative sample of household expenses.

The results, discussed in the following sections, highlight the model's potential to significantly enhance classification accuracy. Additionally, an application will be developed that integrates OCR and classification models or an ensemble learning model, enabling precise categorization of expense empowerment suitable for future application development.

2 Literature Reviews

Optical Character Recognition (OCR) and its applications have seen significant advancements, particularly in extracting data from printed receipts. This review explores the current state of OCR technology, information extraction techniques, product classification models, and ensemble learning approaches in receipt analysis and family expense management.

Ashlin Deepa et al. (2024) [5] presented an automated invoice processing method using Tesseract OCR. The process begins with image preprocessing, including noise reduction, resizing, and contrast enhancement. Text line extraction follows, involving text localization and segmentation. Feature extraction captures visual details used in the embedded text. Convolutional Neural Networks (CNNs) extract visual features, while Long Short-Term Memory Networks (LSTMs) enhance context and word relationships for improved accuracy. The workflow includes uploading and validating the invoice, extracting text with Tesseract, matching it against templates, and structuring the output in JSON format.

Sayallar et al. (2023) [3] developed an advanced OCR engine for printed receipts using deep learning with an open-source solution, Nacsoft OCR, aimed to surpass the performance of established engines like Tesseract and Google Vision API. The methodology employed a three-phase approach: initial processing, word identification, and text recognition, utilizing Convolutional Recurrent Neural Networks (C-RNN) for text interpretation. While this study advanced OCR technology for receipt processing, the inability to share the Turkish dataset used in the evaluation constrained reproducibility. Expanding the scope of OCR applications, Kumar et al. (2020) [6] addressed inefficiencies in the manual processing of paper bills and receipts by developing an application that automated information extraction from bill images. Their approach combined image preprocessing techniques with advanced OCR methods. OpenCV was used for initial image enhancement, including shadow and watermark removal. The Tesseract OCR engine was employed for text extraction, supported by Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks to improve recognition accuracy. The system demonstrated high accuracy for small bills (97.00%) but reduced performance for larger bills (83.00%), likely due to challenges with smaller font sizes. Building on this work, Ha and Horák (2022) [7] showed the OCRMiner system to extract information from invoices in various formats. Their approach emulated human reading by integrating text analysis and layout features. The system combined text analysis techniques with positional layout features and achieved high accuracy rates (90.00% for English, 88.00% for Czech). It demonstrated flexibility in processing various invoice formats, though its reliance on a small training set may limit generalizability.

While not directly related to receipt processing, sentiment analysis and review classification techniques could be valuable for analyzing user feedback on expenses or evaluating the reliability of expense-related information. Alghazzawi et al. (2023) [1] proposed an Ensemble Random Forest-based XGBoost model for precise binary sentiment classification. The approach was tested on two IMDB datasets and the ChnSentiCorp dataset, incorporating preprocessing steps like tokenization, lemmatization, and stemming. The model demonstrated superior accuracy, achieving 98.70% for ChnSentiCorp and 98.20% for IMDB, outperforming existing methods.

Fayaz et al. (2020) [10] presented an ensemble machine-learning model for detecting spam product reviews in a related study. The model integrated the predictions of a Multilayer Perceptron (MLP), K-Nearest Neighbors (KNN), and Random Forest (RF) to classify reviews as either spam or non-spam. The findings demonstrated that the proposed ensemble model significantly outperformed individual classifiers and advanced boosting methods regarding classification accuracy. However, a fundamental limitation was its reliance on the Yelp Dataset, with no evaluation conducted on other datasets to assess the model's generalizability.

Bringing together various aspects of OCR, classification, and analysis, some researchers have focused on developing comprehensive systems for receipt analysis. Huang et al. (2019) [4] focused on automating the analysis of scanned receipts, a critical component for family expense management systems. They established a standardized dataset and evaluation framework for OCR and critical information extraction from scanned receipts. The methodology included tasks like text localization, OCR, and key information extraction, which were evaluated using mean average precision (mAP) and F1 score. Although the competition provided valuable resources and fostered innovation, OCR accuracy still fell short of the 99.00% required for many commercial applications. Nuankaew [11] explores text mining and tokenization techniques, including TF-IDF, Term Frequency, and Binary Term Occurrences, to evaluate job performance in Thai using classification models by breaking down text data into tokens for detailed analysis.

The reviewed literature highlights advancements in OCR, information extraction, and classification models for family expense management systems. OCR accuracy for receipt processing has achieved 90.00%, but challenges persist with diverse receipt formats and physical damage. While product classification models point to potential, most research has focused on manufacturing rather than consumer purchases. Adapting these models for receipt data could improve expense categorization. Addressing these challenges could lead to more accurate and efficient family expense management systems using OCR and machine learning technologies.

3 Methodology

Figure 1 depicts a framework for Leveraging OCR-Driven Information Extraction for Accurate Product Type Classification from Thai Receipt Data: An Ensemble Learning Approach. It begins with converting receipt images into machine-readable text via OCR. The text is then preprocessed and tokenized, showing examples of Thai product names. Following this, feature extraction prepares the data for machine learning models. The processed text is input into a base classification and more complex ensemble models. Finally, the models' performances are evaluated, showcasing the transformation from raw receipt images to accurate product classification.

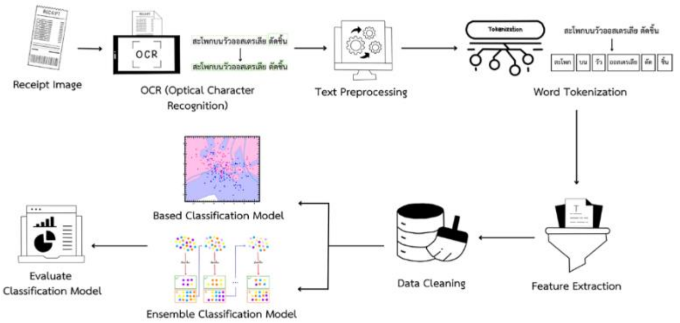
Receipt Image
OCR
สะโพกบนวัวออสเตรเลีย ตัดชิ้น
สะโพกบนวัวออสเตรเลีย ตัดชิ้น
OCR (Optical Character Recognition)
Text Preprocessing
Tokenization
สะโพกบนวัวออสเตรเลีย ตัดชิ้น
สะโพก
บน
วัว
ออสเตรเลีย
ตัด
ชิ้น
Word Tokenization
Based Classification Model
Ensemble Classification Model
Evaluate Classification Model
Data Cleaning
Feature Extraction

Fig. 1. Research Framework

3.1 Data Collection

The dataset consists of 1,305 images of electronic receipts, printed receipts, and PDF files, contributed by 100 volunteers and receipt data donors in Phayao Province, Thailand, during 2023-2024.

Fig. 2. Examples of purchase receipts

3.2 Information Extraction Process

Purchase receipts are official documents confirming payment for goods or services. They typically include Payer's and payee's names, Transaction amount, Item details (product name, brand, model, quantity), Receipt or tax invoice number, Tax ID, Names and addresses of both parties, VAT amount etc.

The text extraction process from images utilized the Tesseract OCR engine [12] for the Thai language. This open-source tool, developed by Google, was specifically focused on recognizing product names. Product category labels were determined based on the department store's classification on the purchase receipt. For examples include “ปูม้าตัวเมียแช่แข็ง” (Food), “กบติดหนัง” (Food), “สำลีอนามัย 50 ซอง” (Health & Beauty), “ผ้าพันแผล 2 นิ้ว 5 แผ่น 20 ถุง” (Health & Beauty), “ถุงซักผ้าพิมพ์ลายขนาด 40 x 50 ซม.” (Home & Lifestyle), “กล่องอเนกประสงค์ 50 ล.” (Home & Lifestyle), ““ปืนจุดเตาแก๊ส” (Household), “แก้วชงชากาแฟ แบบกด 600 มล.” (Household) and others. The data shown in Table 1. Thai OCR presents unique challenges due to the complexity of its script and the lack of word boundaries. To address these, researchers applied Thai-specific text normalization following the OCR process.

Table 1. Distribution of Products Across Categories

Product Type	Number of products	Product Type	Number of products
Drinks	381	Home & Lifestyle	611
Food	2,021	Household	1,366
Health & Beauty	708	**Total:**	**5,087**

Table 1 provides an overview of the distribution of products across different categories. The dataset comprises 5,087 products across five categories, with Food being the largest category (2,021 products) and Drinks the smallest (381 products).

3.3 Preprocessing Process

The researcher implemented a comprehensive data preprocessing strategy. This process began with thorough data cleansing, removing extraneous information, and reducing noise and incomplete data to enhance data quality. For word tokenization, the PyThaiNLP library was employed with the 'newmm' engine, which utilizes a maximal matching algorithm specifically designed for the nuances of the Thai language. This method efficiently segmented continuous text into meaningful word units, forming a critical foundation for subsequent feature extraction and classification processes [13] [14] (https://pythainlp.org/docs/2.1/api/tokenize.html).

Applying TF-IDF to Thai text required careful consideration of language-specific characteristics. After tokenization, term frequency (TF) was calculated for each word within a product name, and inverse document frequency (IDF) was determined across the entire corpus of product names. To address the challenges inherent in Thai, character-level n-grams were incorporated alongside word-level features to capture sub-word information, which is essential for managing the language's long compound words [15]. Stopword removal was also conducted using a custom Thai stopword list, allowing for the effective representation of word significance in the context of Thai product names and balancing common terms with unique features specific to each product category.

After removing rows where the conversion from string to float failed, 5,072 rows of data remained. This refined dataset, with well-defined features, provided a robust foundation for subsequent model training and evaluation.

3.4 Classification Model

In this study, the researchers selected a diverse array of classification models to construct an effective system. These models can be categorized into two main groups:

Base classifiers , including K-Nearest Neighbors, Logistic Regression, Naive Bayes, Decision Trees, and Support Vector Machines, were selected for their diverse approaches to classification, each offering unique strengths in handling text data. KNN [16] was chosen for its effectiveness with local patterns, Logistic Regression for its interpretability, Naive Bayes for its efficiency with text [11], Decision Trees for capturing non-linear relationships, and SVM [11] for its robustness in high-dimensional spaces.

For ensemble methods, Random Forest [17], AdaBoost, Extra Trees, Bagging [16], and Majority Voting were employed to leverage the strengths of multiple models while mitigating individual weaknesses. Random Forest and Extra Trees were selected for their capability to handle high-dimensional data and reduce overfitting. AdaBoost was included for its focus on hard-to-classify instances. Bagging was used to reduce variance, and Majority Voting was applied to combine diverse model predictions [18].

Ensemble methods were constructed using scikit-learn's implementations. Cross-validation was utilized to optimize hyperparameters for each base model and ensemble. The training process involved fitting each base model independently, followed by combining them according to the specific algorithm of the ensemble method.

The research aims to capture different aspects of the data, overcome limitations of individual models, and ultimately construct a more robust and effective classification system. This combination of methods allows for a comprehensive analysis of the dataset, potentially leading to more accurate and reliable results.

3.5 Data Split and Model Training

The product names dataset of 5,072 samples from Thai receipts was split into training and testing sets using a 70:30 ratio. This resulted in 3,550 samples for training and 1,522 for testing. The training set was used to teach classification models to recognize patterns in OCR-extracted text and features. The testing set served to evaluate the models' performance on unseen data, simulating real-world scenarios. This approach ensures robust assessment of the models' ability to generalize.

3.6 Evaluation

The research evaluated classification models using key performance indicators, including accuracy, precision, recall, F1-score, Macro average, and Weighted average to assess their performance.

4 Research Results

This research evaluated various classification algorithms, including both base classifiers and ensemble methods, to determine their effectiveness in product classification from product name using OCR-extracted data from receipts. The results demonstrate the superiority of ensemble learning techniques over individual classifiers in most scenarios. Presentation of model performance for Base Classification and Ensemble Learning In Table 2 and Table 3.

Table 2. Result of Base Classification Algorithms

Class Label	**KNN**			**Logistic Regression**		
	Accuracy: 59.00			**Accuracy: 90.28**		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Drinks	96.55	22.58	36.60	100.00	59.68	74.75
Food	50.16	99.84	66.77	88.24	98.24	92.97
Health & Beauty	98.57	34.50	51.11	93.01	86.50	89.64
Home & Lifestyle	100.00	38.17	55.25	97.96	77.42	86.49
Household	99.06	27.20	42.68	88.04	95.34	91.54
Macro Avg.	88.87	44.46	50.48	93.45	83.43	87.08
Weighted Avg.	78.79	59.00	54.74	90.96	90.28	89.89
Class Label	**Naive Bayes**			**Decision Tree**		

Class Label	**KNN**			**Logistic Regression**		
	Accuracy: 59.00			**Accuracy: 90.28**		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score
	Accuracy: 86.60			**Accuracy: 83.84**		
	Precision	Recall	Precision	Recall	Precision	Recall
Drinks	100.00	25.00	40.00	80.85	61.29	69.73
Food	84.52	98.56	91.00	87.04	91.21	89.08
Health & Beauty	92.31	84.00	87.96	78.85	82.00	80.39
Home & Lifestyle	97.78	70.97	82.24	86.54	72.58	78.95
Household	83.33	95.85	89.16	80.88	85.49	83.12
Macro Avg.	91.59	74.88	78.07	82.83	78.52	80.25
Weighted Avg.	88.12	86.60	84.91	83.84	83.84	83.61
Class Label	**SVM**					
	Accuracy: 92.51					
	Precision	Recall	Precision			
Drinks	98.78	65.32	78.64			
Food	90.44	98.24	94.18			
Health & Beauty	94.82	91.5	93.13			
Home & Lifestyle	98.73	83.33	90.38			
Household	91.22	96.89	93.97			
Macro Avg.	94.80	87.06	90.06			
Weighted Avg.	92.91	92.51	92.26			

Table 2 provides a comprehensive comparison of five distinct classification algorithms. Initially, SVM performed the most balanced and robust performance across all categories, with a weighted average F1-score of 92.26% and accuracy of 92.51%. It excelled particularly in the Drinks category (98.78% precision, 65.32% recall) and maintained consistent performance across other categories.

In contrast, Logistic Regression demonstrated high precision for Drinks (100%) and Food (98.24%) categories but struggled with recall in some classes. Meanwhile, the Decision Tree algorithm exhibited strong performance in Health & Beauty and Home & Lifestyle categories, while Naive Bayes performed exceptionally well in the Food category (91% precision, 87.04% recall). KNN, despite high precision in some categories, suffered from poor recall, particularly in the Drinks (22.58%) and Household (27.20%) categories, indicating potential overfitting issues.

Table 3. Result of Ensemble Learning Algorithms

Class Label	**Random Forest**			**AdaBoost**		
	Accuracy: 90.54			**Accuracy: 64.85**		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Drinks	96.30	62.90	76.10	71.43	16.13	26.32
Food	88.29	98.72	93.21	62.80	96.01	75.93
Health & Beauty	90.26	88.00	89.11	68.18	60.00	63.83
Home & Lifestyle	96.25	82.80	89.02	88.98	56.45	69.08
Household	91.19	91.19	91.19	58.03	36.53	44.83
Macro Avg.	92.46	84.72	87.73	69.88	53.02	56.00
Weighted Avg.	90.91	90.54	90.25	66.20	64.85	61.58

Class Label	**Random Forest**			**AdaBoost**		
	Accuracy: 90.54			**Accuracy: 64.85**		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Class Label	**Extra Trees**			**Bagging**		
	Accuracy: 92.05			**Accuracy: 87.45**		
	Precision	Recall	Precision	Recall	Precision	Recall
Drinks	95.70	71.77	82.03	89.66	62.90	73.93
Food	91.10	98.08	94.46	87.28	96.49	91.65
Health & Beauty	92.39	91.00	91.69	82.27	83.50	82.88
Home & Lifestyle	96.36	85.48	90.60	94.81	78.50	85.88
Household	90.84	92.49	91.66	87.05	87.05	87.05
Macro Avg.	93.28	87.77	90.09	88.21	81.69	84.28
Weighted Avg.	92.22	92.05	91.90	87.68	87.45	87.18
Class Label	**Majority Voting**					
	Accuracy: 91.92					
	Precision	Recall	Precision			
Drinks	96.74	71.77	82.41			
Food	90.13	99.20	94.45			
Health & Beauty	92.15	88.00	90.03			
Home & Lifestyle	97.48	83.33	89.86			
Household	91.56	92.75	92.15			
Macro Avg.	93.61	87.01	89.78			
Weighted Avg.	92.20	91.92	91.74			

Table 3 presents the performance metrics of various ensemble learning algorithms. Majority Voting emerged as the top performer among the algorithms, demonstrating balanced and robust results across all categories. It achieved a weighted average F1-score of 91.74% and an accuracy of 91.92%, with notable success in the Drinks category. Similarly, Extra Trees recorded the highest overall accuracy at 92.05%, excelling in the Food, Health & Beauty categories. Random Forest also showed strong performance, particularly in the Food category, with a weighted average F1 score of 90.25% and an accuracy of 90.54%.

Bagging delivered solid results, achieving a weighted average F1-score of 87.18% and an accuracy of 87.45%. In contrast, AdaBoost underperformed compared to the other algorithms, recording lower scores across most categories.

5 Discussion

The varied performance of algorithms underscores the complexity of classifying product types from OCR-extracted receipt data based on product names. SVM's strong results are due to its ability to handle complex decision boundaries and robustness in high-dimensional spaces, especially with text-based features. Logistic Regression's high precision but lower recall, as seen in the Drinks category, suggests possible conservatism due to class imbalance. The Decision Tree excels in capturing category-specific features but shows variability across product types. Naive Bayes performs well in the Food category but struggles with diverse features. KNN's high precision but poor recall indicates overfitting, difficulty generalizing, and familiarity with high-dimensional or imbalanced data.

The significant potential of Ensemble methods, particularly Majority Voting and Extra Trees, excelled in classifying product types from OCR-extracted names. Their success stems from combining multiple weak learners, improving generalization, and reducing overfitting. Majority Voting demonstrated robust performance across categories, especially in Drinks, indicating its ability to handle diverse product names and potential OCR errors. Meanwhile, Extra Trees outperformed in Food and Health & Beauty categories, suggesting superior discernment of subtle differences in product names. In addition, Random Forest showed strong recall in the Food category, likely due to distinctive naming patterns. Its effectiveness in identifying food products was notable.

Conversely, AdaBoost underperformed, particularly in the Drinks and Household categories. It may indicate challenges with class imbalance or specific product naming characteristics, highlighting areas for potential improvement in future iterations. OCR errors in mixed Thai-English product names can cause misclassification. Challenges arise from missing text due to faded stamps, faint ink, or damaged receipts. These issues underscore the need for robust preprocessing and error-handling mechanisms to enhance classification accuracy when working with imperfect OCR output from receipts.

6 Conclusion

This research demonstrates the significant potential of leveraging OCR-driven information extraction for accurate product type classification from receipts based on product names. Ensembled algorithms methods, particularly Majority Voting, and Extra Trees algorithms emerged as the most effective approaches, offering a balance of precision and recall across diverse product categories. These methods prove well-suited for handling the challenges associated with OCR-extracted product names, including potential errors and variations in naming conventions.

The study's findings pave the way for more sophisticated and accurate product classification systems in retail analytics and expense management applications. However, the varied performance across algorithms and categories underscores the complexity of the task and the need for further research.

Future work should focus on several key areas to enhance classification accuracy and robustness. Advanced feature extraction techniques specific to product names, such as incorporating word embeddings or pre-trained language models, could improve performance. Investigating the impact of OCR quality on classification and developing methods to handle OCR errors is crucial. Addressing class imbalance issues and incorporating additional contextual information from receipts may further boost accuracy. Exploring deep learning approaches like recurrent neural networks or transformers could more effectively capture sequential patterns in product names. For Thai-specific applications, developing language-specific models that consider the unique characteristics of Thai in OCR and text processing is essential.

In conclusion, while this study yields promising results in OCR-based product type classification using ensemble methods, it also highlights the need for continued research to overcome challenges in handling diverse product naming conventions and

OCR-specific data characteristics. The success of ensemble methods sets a strong foundation for future advancements in this field, potentially revolutionizing retail analytics and into real-time OCR and classification systems for family expense management applications. The data collection for this research was conducted carefully to protect personal information. Names, addresses, and other identifying details were removed from the receipts before processing to ensure the privacy of the data donors.

7 Conflict of Interest

The researchers declare that there is no conflict of interest for this research.

8 Acknowledgement

This research project was supported by the Thailand Science Research and Innovation Fund and the University of Phayao. It also received support from many advisors, academics, researchers, students, and staff. The authors thank everyone for their support and cooperation in completing this research.

Moreover, the special thanks to Claude, ChatGPT, Sci Space, Gemini, and Perplexity Generative AI for their invaluable assistance in gathering information, reading, and providing guidance throughout this research.

References

1. Alghazzawi, D.M., Alquraishee, A.G.A., Badri, S.K., Hasan, S.H.: ERF-XGB: Ensemble Random Forest-Based XG Boost for Accurate Prediction and Classification of E-Commerce Product Review. Sustainability. 15, 7076 (2023). https://doi.org/10.3390/su15097076.
2. Saout, T., Lardeux, F., Saubion, F.: An Overview of Data Extraction From Invoices. IEEE Access. 12, 19872–19886 (2024). https://doi.org/10.1109/ACCESS.2024.3360528.
3. Sayallar, C., Sayar, A., Babalik, N.: An OCR Engine for Printed Receipt Images using Deep Learning Techniques. Int. J. Adv. Comput. Sci. Appl. IJACSA. 14, (2023). https://doi.org/10.14569/IJACSA.2023.0140295.
4. Huang, Z., Chen, K., He, J., Bai, X., Karatzas, D., Lu, S., Jawahar, C.V.: ICDAR2019 Competition on Scanned Receipt OCR and Information Extraction. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 1516–1520 (2019). https://doi.org/10.1109/ICDAR.2019.00244.
5. R N, A.D., Chinta, S., Ashili, N.K., Babu, B.S., Vydugula, R.R., VSL, R.S.: An Intelligent Invoice Processing System Using Tesseract OCR. In: 2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS). pp. 1–6 (2024). https://doi.org/10.1109/ADICS58448.2024.10533509.
6. Kumar, V., Kaware, P., Singh, P., Sonkusare, R., Kumar, S.: Extraction of information from bill receipts using optical character recognition. 2020 Int. Conf. Smart Electron. Commun. ICOSEC. 72–77 (2020). https://doi.org/10.1109/ICOSEC49089.2020.9215246.

7. Ha, H.T., Horák, A.: Information extraction from scanned invoice images using text analysis and layout features. Signal Process. Image Commun. 102, 116601 (2022). https://doi.org/10.1016/j.image.2021.116601.
8. Yindumathi, K.M., Chaudhari, S.S., Aparna, R.: Analysis of Image Classification for Text Extraction from Bills and Invoices. In: 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT). pp. 1–6 (2020). https://doi.org/10.1109/ICCCNT49239.2020.9225564.
9. Gan, B., Zhang, C.: An Improved Model of Product Classification Feature Extraction and Recognition Based on Intelligent Image Recognition. Comput. Intell. Neurosci. 2022, 2926669 (2022). https://doi.org/10.1155/2022/2926669.
10. Fayaz, M., Khan, A., Rahman, J.U., Alharbi, A., Uddin, M.I., Alouffi, B.: Ensemble Machine Learning Model for Classification of Spam Product Reviews. Complexity. 2020, 8857570 (2020). https://doi.org/10.1155/2020/8857570.
11. Nuankaew, W., Thipmontha, R., Jeefoo, P., Nasa-ngium, P., Nuankaew, P.: Using Text Mining and Tokenization Analysis to Identify Job Performance for Human Resource Management at the University of Phayao. Presented at the September 29 (2023). https://doi.org/10.1007/978-3-031-42430-4_47.
12. Adjetey, C., Adu-Manu, K.S.: Content-based Image Retrieval using Tesseract OCR Engine and Levenshtein Algorithm. Int. J. Adv. Comput. Sci. Appl. IJACSA. 12, (2021). https://doi.org/10.14569/IJACSA.2021.0120776.
13. Pythainlp: pythainlp.tokenize — PyThaiNLP <unknown> documentation, https://pythainlp.org/docs/2.1/api/tokenize.html#pythainlp-tokenize, last accessed 2024/08/13.
14. Kongsumran, N.: Thai tokenizer invariant classification based on bi-lstm and distilbert encoders. Chulalongkorn Univ. Theses Diss. Chula ETD. (2021). https://doi.org/10.58837/CHULA.THE.2021.113.
15. Mohammed, M.T., Rashid, O.F.: Document retrieval using term term frequency inverse sentence frequency weighting scheme. Indones. J. Electr. Eng. Comput. Sci. 31, 1478–1485 (2023). https://doi.org/10.11591/ijeecs.v31.i3.pp1478-1485.
16. Chicho, B.T., Abdulazeez, A.M., Zeebaree, D.Q., Zebari, D.A.: Machine Learning Classifiers Based Classification For IRIS Recognition. Qubahan Acad. J. 1, 106–118 (2021). https://doi.org/10.48161/qaj.v1n2a48.
17. Adam, H., Muhammad, A., Aboaba, A.A.: Design of a Hybrid Machine Learning Base-Classifiers for Software Defect Prediction. Int. J. Innov. Res. Dev. (2022). https://doi.org/10.24940/ijird/2022/v11/i10/OCT22020.
18. Nuankaew, W.S., Bussaman, S., Nuankaew, P.: Evolutionary Feature Weighting Optimization and Majority Voting Ensemble Learning for Curriculum Recommendation in the Higher Education. In: Surinta, O. and Kam Fung Yuen, K. (eds.) Multi-disciplinary Trends in Artificial Intelligence. pp. 14–25. Springer International Publishing, Cham (2022). https://doi.org/10.1007/978-3-031-20992-5_2.