

Design and Development of a Web-Based Tax Invoice Recapitulation Application Using the Optical Character Recognition (OCR) Method

Zulfi Andriansah^{1✉}, Arip Kristiyanto²

^{1,2} Department of Information System, Faculty of Computer Science, Universitas Pamulang, Indonesia

✉ Corresponding Author : Zulfi Andriansah (e-mail: zulfiandriansah23@gmail.com)

Article Information	ABSTRACT
Article History Received : August 23, 2025 Revised : September 16, 2025 Published : October 04, 2025	The process of recapitulating tax invoice data is crucial because it directly relates to a company's financial management. This process can be challenging for business owners when implemented using inefficient methods. For example, at CV Banten Inspirasi Mandiri, tax invoice data recapitulation is still carried out conventionally by manually entering data into Microsoft Excel. This approach is ineffective in terms of time efficiency and offers limited guarantees of data accuracy. To address this issue, an Artificial Intelligence (AI)-based application was developed to optimize the company's time and resources. Specifically, this system utilizes Optical Character Recognition (OCR) technology, which enables automatic data extraction from tax invoice files. The implementation results showed a time efficiency rate of 92.92%, with 100% accuracy for OCR results from PDF files and 89% for OCR results from JPG files, outperforming conventional methods, which only achieved 80% accuracy.
Keywords: <i>Artificial Intelligent; Optical Character Recognition; Recapitulation; Tax; Tax Invoice.</i>	

INTRODUCTION

In a company, tax is an obligation that must be fulfilled by the business entity in order to fulfill its obligations to the state, one of the obligations in fulfilling the company's tax needs is the Company's financial tax reporting or notification letter which is carried out every month and year to the Direktorat Jendral Pajak (DJP) of the Republic of Indonesia as stated in Law No. 7 of 2021 (Anggraeni et al., 2025), this will of course go through a process of recapitulating tax invoices that have been received and issued by the company in conducting transactions (Hartanto Dharmawan, 2023). The process of managing data listed on tax invoices is one of the important things in company operations, In this process, of course, it can be one of the obstacles experienced by business entities because the process is quite complicated if done using conventional methods.

This happened to CV. Banten Inspirasi Mandiri, this company is engaged in the procurement of goods and services, where in every transaction carried out it will receive or issue a tax invoice, of course this business entity is required to report tax financial reports every month and year, one of the things that must be done is to recapitulate the tax invoices that have been received and issued. However, after conducting interviews and observations with employees of CV. Banten Inspirasi Mandiri Company, there are obstacles in processing tax invoice data, namely the recapitulation process which is still carried out using conventional methods, by manually entering one by one data into the Microsoft Excel application, of course this method is less effective in terms of time efficiency (Alexander & Pujawan, 2024), and the level of data acquisition resulting from the process is less often experiencing errors during data recapitulation (Muhammad Yahya et al., 2023), along with the development of technology, these problems can be solved by creating

an Artificial Intelligent (AI) application program that has the ability to spend time and Company resource. And also the recapitulation process is carried out systematically to minimize data preparation errors (Nugraha, 2024). In this case, the artificial intelligence technology used is Optical Character Recognition (OCR) (Wahyuddin & Hasim, 2023) which can retrieve data contained in tax invoice files and extract it to be processed into a company's tax financial report (Reyvansyah, 2023).

Previous research using optical character recognition technology for tax invoice data extraction has yielded positive results, demonstrating the ability to extract required information from tax invoice documents with a higher degree of accuracy. Thus, this demonstrates that Optical Character Recognition (OCR) technology is suitable for use in tax invoice data recapitulation systems (Rosnita & Adek, 2023), with the need for several feature and technology optimizations.

Previous studies have not integrated OCR technology with field mapping, data parsing, and data cleaning. The distinction of this study lies in the design of the developed application system, which features automatic tax invoice data recapitulation using Optical Character Recognition (OCR) (Almash & Indriyanti, 2022). In addition, this study introduces technological features not explored in earlier research, such as a data parsing function that groups OCR-generated text into multiple fields based on predefined patterns (Nashiruddin et al., 2025). The field mapping capability enables the parsed data to be organized into columnar form. Furthermore, the system incorporates a data cleaning function that removes invalid or irrelevant characters or data from the OCR results (Sukarna et al., 2025). This tax invoice recapitulation system is built using the Python programming language (Rahman et al., 2023) integrated with the OCR library Tesseract OCR and PDFplumber (Kusnanto et al., 2022). The system also uses MySQL Server as a database management system to store and manage extracted tax invoice data. This website-based application system makes it easy for users to use (Andriansah et al., 2025).

RESEARCH METHODS

A research method is a series of steps or stages carried out, starting from problem analysis to conclusions. This study used the research and development (RnD) method. Figure 1 illustrates the stages of the workflow used in this study.

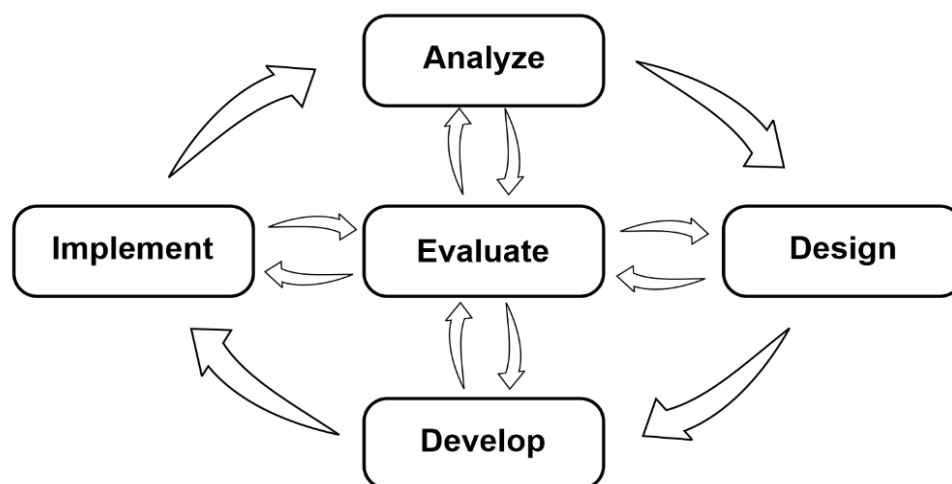


Figure 1. Research and Development Methods

Application of the ADDIE model (Analysis, Design, Development, Implementation, Evaluation), with a description of the activities carried out as follows.

Table 1. Development Stages

Stages	Activities
Analysis	analyze the problems that occur and determine the needs needed to resolve the problems that arise.
Design	system design as a solution to problems, starting from creating concepts, specifications and also the workflow of the system to be developed.
Development	The previously created system design is implemented and further developed using the Python programming language, with Tesseract OCR (Meirza & Puteri, 2024) and PDFplumber modules utilized for data extraction (Iskandar & Kurniawati, 2025).
Implementation	testing a system that has been successfully developed, validating whether the system created is running according to the development objectives or whether there is something that needs to be evaluated.
Evaluation	System testing results were evaluated using black-box testing to ensure that the application functioned as intended. In addition, time efficiency was measured using the time-and-motion method, while data accuracy was assessed through the confusion matrix method by comparing the OCR output with the actual data.

RESULTS AND DISCUSSION

Results

Analysis

Based on direct interviews with CV. Banten Inspirasi Mandiri, several issues were identified:

- The tax invoice data recapitulation process still uses conventional methods, requiring manual data entry.
- The large number of tax invoices causes the tax invoice recapitulation process to be time-consuming, and administrators also experience difficulties managing the data.
- Data recapitulation errors often occur due to administrative errors during data collection.
- It is difficult to check monthly and annual tax invoice data.

The solution to this problem is to develop a web-based tax invoice recapitulation application using optical character recognition (OCR). This can replace and optimize existing conventional methods, thereby increasing processing time efficiency and data accuracy (Khansyah & Khotimah, 2025).

Design

The system design is based on needs analyzed through observations and interviews at CV. Banten Inspirasi Mandiri. To illustrate the system's workflow and structure, the following system design was created:

1. Use Case Diagram

In system design, a use case diagram is required to illustrate the relationships between actors and the system. A use case diagram contains actor and system elements that can be connected using an include relationship, which states that the functionality of the base use case can only be fulfilled with the assistance of the supplier use case. The extend relationship is also used to indicate that the functionality of the base use case can

be extended by the supplier. The following is a use case created in the design of a tax invoice recapitulation system.

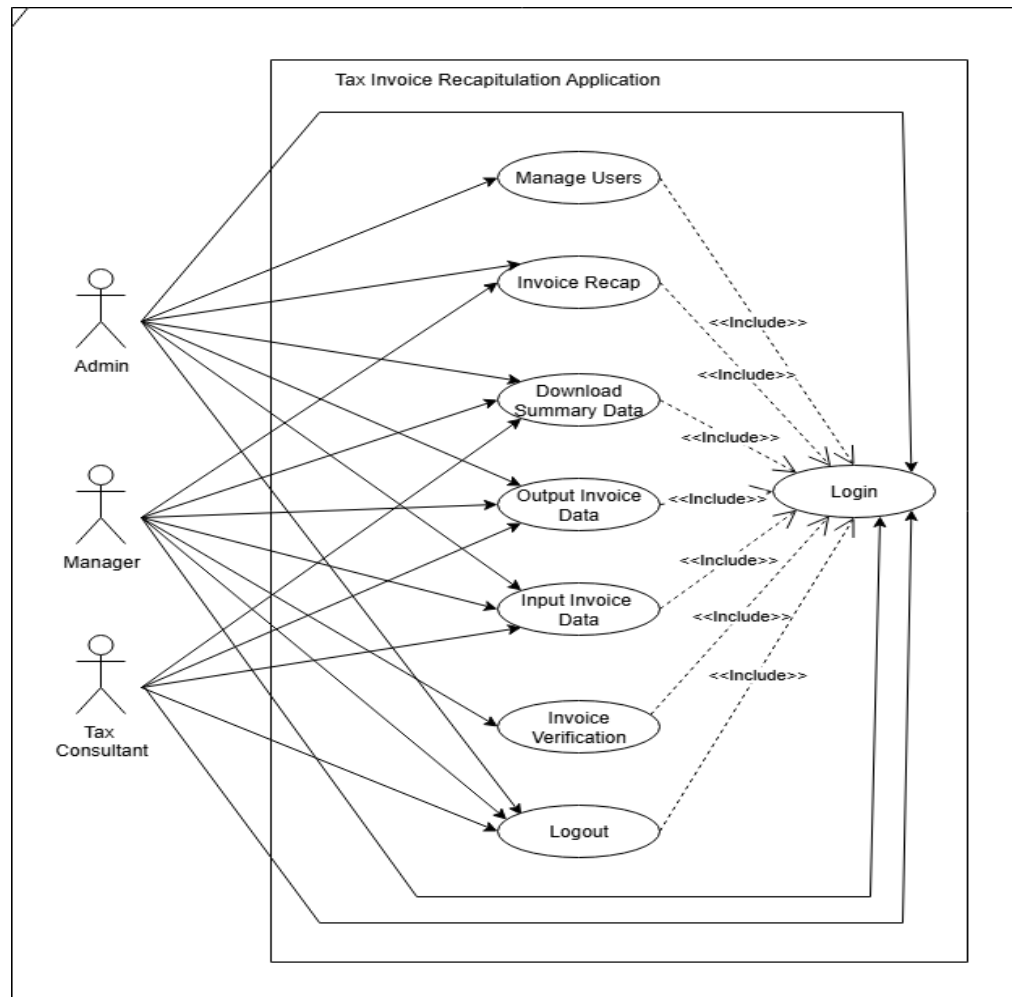


Figure 2. Use Case Diagram

Figure 2 depicts a use case diagram for the tax invoice recapitulation program. The diagram identifies three actors or users who will use this application: the admin, the manager, and the tax consultant. Each actor has different access rights. For example, the admin has access to manage user data, summarize invoices, download invoice data, access output invoice data, and access input invoice data. Similarly, the manager has access to summarize invoices, download invoice data, access output invoice data, access input invoice data, and verify invoice data. Finally, the tax consultant has access to download invoice data, access output invoice data, and access input invoice data. All actors must verify application access through the login process for application security, and log out when not using the application.

2. System Flow

The following illustrates the system workflow of the tax invoice recapitulation application using the Optical Character Recognition (OCR) method.

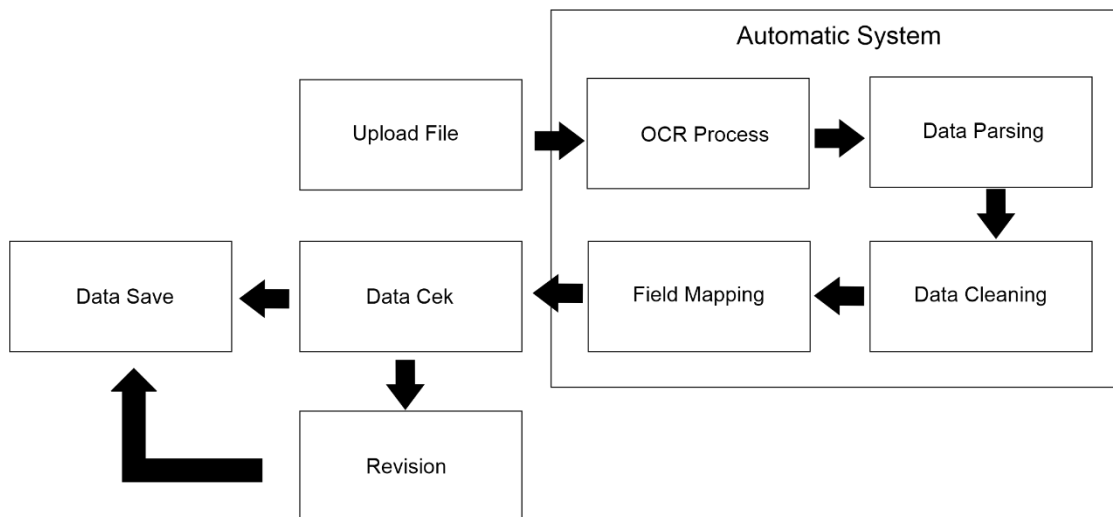


Figure 3. System Flow

3. Interface Design

The following is an interface design for a tax invoice recapitulation system.

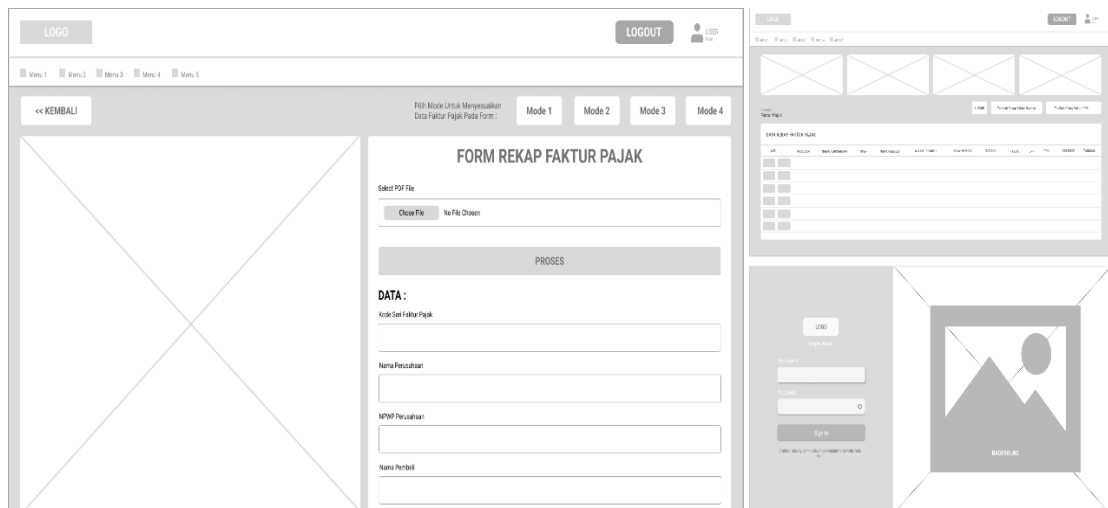


Figure 4. Wireframe Interface Design

Development

Interface Implementation

Based on the interface design of the tax invoice recapitulation application system, the following are the results of implementing the design on each page of the application.

1. Login Page

On the login page, there is a form that allows users to enter their account credentials when logging into the application.

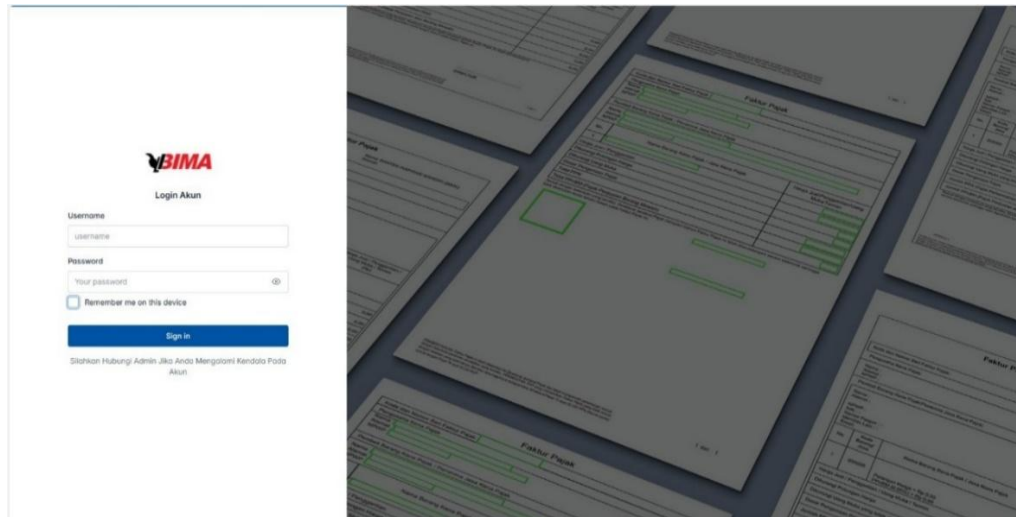


Figure 5. Login Page

2. Dashboard Page

On the admin dashboard page, various information is displayed, including the page menu accessible to the admin, financial summary cards, summary data, and buttons for adding or downloading the summary data.

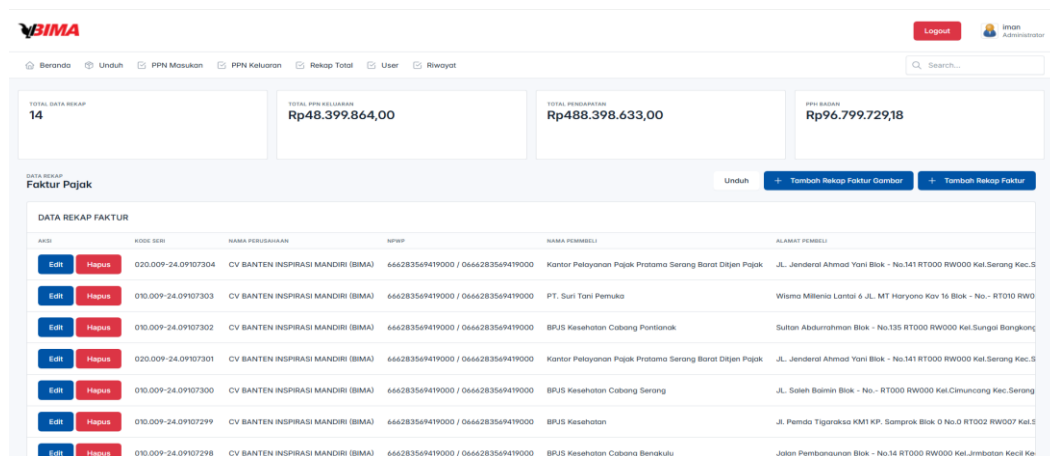


Figure 6. Dashboard Page

3. Recapitulation Page

On the tax invoice recap page, several menu options are available that can be selected according to the tax invoice file, along with a Back button, a file upload form, and the data extraction results.

Figure 7. Recapitulation Page

Evaluation

At this stage, the system will be tested using the black-box testing method. This method focuses on evaluating the functionality of the developed system (Nurwicaksono et al., 2023). The results obtained will serve as input for system evaluation. The following are the test results for the developed mail management system.

Table 2. Testing Results

Test Code	User Needs	Test Results
TC001	Admin, Manager, and Tax Consultant users can log in using their registered accounts.	Succeed
TC002	Admin and Manager users can perform tax invoice recapitulation using the Optical Character Recognition (OCR) method.	Succeed
TC003	Admin, Manager, and Tax Consultant users can download the summary data.	Succeed
TC004	Admin, Manager, and Tax Consultant users can access the summary data.	Succeed
TC005	Admin users can manage other users by adding or deleting user accounts	Succeed
TC006	Manager users can verify the validity of invoices to determine whether they are valid or invalid.	Succeed
TC007	Admin, Manager, and Tax Consultant users can log out of the system.	Succeed

Discussion

System Efficiency Calculation

The efficiency of this system is tested using a data calculation method called Time and Motion. This method measures system efficiency based on the duration of an activity cycle (Annisawati et al., 2021). In this test, five sample invoice files are used to evaluate the speed of the recapitulation process.

Table 3. OCR Method Time Result

Test Result	Work Element (/ Second)			
	Upload Invoice	OCR Process	Data Cek	Data Save
Tax Invoice File 1	04,21	04,10	05,93	03,03

Tax Invoice File 2	04,41	03,58	04,43	04,44
Tax Invoice File 3	05,27	02,76	05,51	03,24
Tax Invoice File 4	04,38	02,68	05,72	03,24
Tax Invoice File 5	04,40	04,82	03,38	03,37
Avarage	4,534	3,588	4,994	3,554
Total	16,670			

Table 4. Convensional Method Time Result

Test Result	Work Element (/ Second)			
	Select Invoice	Data Input	Data Cek	Data Save
Tax Invoice File 1	03,27	189,39	06,52	13,02
Tax Invoice File 2	03,71	206,30	05,47	02,48
Tax Invoice File 3	04,46	171,18	07,04	03,27
Tax Invoice File 4	03,72	181,36	04,77	01,37
Tax Invoice File 5	04,02	161,76	07,47	02,89
Avarage	3,836	182,00	6,254	4,406
Total	196,496			

Standard Time = Avarage x (1 + Allowence)

Standard Time OCR Method (X) = 16,670 x 1,10 = **18,337 Second**

Standard Time Convensional Method (Y) = 196,496 x 1,15 = **226,970 Second**

$$Efisiency = \frac{Y-X}{Y} \times 100\% \quad Efisiency = \frac{226,970-18,337}{226,970} \times 100\% = \mathbf{91,92\%}$$

System Accuracy Calculation

This data accuracy test employs the confusion matrix calculation method (Ardiningtyas et al., 2021), which compares the actual data with the OCR results. The same process was also applied to calculate the data accuracy level using conventional methods, yielding an accuracy of 80%. This test was conducted using ten sample tax invoice data files, with the resulting data presented as follows.

```

=== DETAIL KESALAHAN PER FIELD ===

Total Kesalahan : 0
Total Kecocokan : 100
Akurasi Per Field: 100.00%

```

Figure 8. PDF File OCR Result Data**Table 5.** Valid and Invalid Data Details of PDF File OCR Method

No	Document	Valid Data Attributes	Invalid Attribute Data
1	Tax Invoice File 1	10	0
2	Tax Invoice File 2	10	0
3	Tax Invoice File 3	10	0
4	Tax Invoice File 4	10	0
5	Tax Invoice File 5	10	0
6	Tax Invoice File 6	10	0
7	Tax Invoice File 7	10	0
8	Tax Invoice File 8	10	0

9	Tax Invoice File 9	10	0
10	Tax Invoice File 10	10	0

The data obtained from the results of data extraction of tax invoice PDF files using the OCR method, then the calculation will be carried out using a confusion matrix, as shown in the following equation.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad Accuracy = \frac{100 + 0}{100 + 0 + 11 + 0} = 1 = \mathbf{100\%}$$

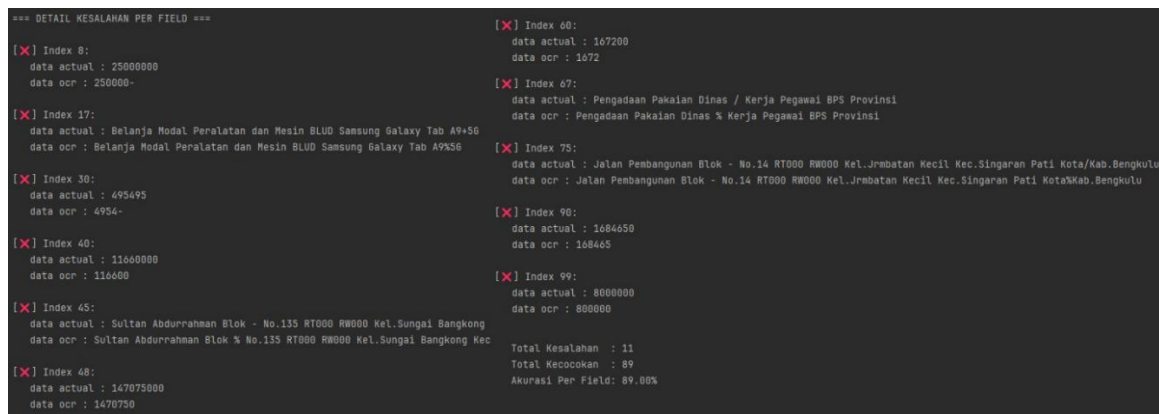


Figure 9. JPG File OCR Result Data

Table 3. Valid and Invalid Data Details of JPG File OCR Method

No	Document	Valid Data Attributes	Invalid Attribute Data
1	Tax Invoice File 1	9	1
2	Tax Invoice File 2	9	1
3	Tax Invoice File 3	9	1
4	Tax Invoice File 4	9	1
5	Tax Invoice File 5	8	2
6	Tax Invoice File 6	9	1
7	Tax Invoice File 7	9	1
8	Tax Invoice File 8	9	1
9	Tax Invoice File 9	9	1
10	Tax Invoice File 10	9	1

The data obtained from extracting tax invoice JPG files using the OCR method contained several errors caused by suboptimal image quality and unclear scan results. These issues led to inaccuracies in several data attributes. To evaluate these errors, calculations were performed using a confusion matrix, as shown in the following equation.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad Accuracy = \frac{89 + 0}{89 + 0 + 11 + 0} = 0.89 = \mathbf{89\%}$$

CONCLUSION

Based on the results of the implementation and testing of the system using the black-box testing method, the application runs well according to the desired design, with a time efficiency level of 92.92% and a data accuracy level of 89%. In comparison, the conventional method produces a data accuracy level of 100% for OCR results from PDF files and 89% for OCR results from JPG files. Overall, the proposed system achieves higher accuracy than the conventional approach which only reaches 80%. Thus, it can be concluded that the tax invoice recapitulation

application system using Optical Character Recognition (OCR) provides a more efficient and accurate solution for the tax invoice data recapitulation process at CV Banten Inspirasi Mandiri.

From an academic perspective, this research provides valuable insights into the application of Optical Character Recognition (OCR) technology in the document recapitulation process. The use of OCR facilitates a more efficient and automated approach to document recapitulation. This study integrates OCR with supporting technologies such as field mapping, data parsing, and data cleaning to ensure that the resulting data conforms to the required format. These contributions enrich the academic discourse on OCR applications by offering a deeper understanding of its role in enhancing operational efficiency and data accuracy. Consequently, this research serves as a reference for future studies and can guide both academics and practitioners in developing improved tax invoice document recapitulation systems.

REFERENCES

- Alexander, D., & Pujawan, I. N. (2024). Perancangan Sistem Warehouse Berbasis Teknologi Ocr Untuk Meningkatkan Efektivitas Dan Efisiensi. *Jurnal Ilmiah Teknik Industri*, 12(1), 1–11. <https://doi.org/10.24912/jitiuntar.v12i1.28100>
- Almash, K. A., & Indriyanti, A. D. (2022). Identifikasi Merek Obat Menggunakan Metode OCR dengan Perbandingan Morphological Opening dan Closing. *Journal of Informatics and Computer Science (JINACS)*, 3(04), 494–505. <https://doi.org/10.26740/jinacs.v3n04.p494-505>
- Andriansah, Z., Robani, R., Qomaruzzaman, J., & Kristiyanto, A. (2025). Pengembangan Aplikasi Manajemen Surat Berbasis Website Menggunakan Metode Waterfall pada PT Bima Karya Reksatama. *Jurnal Teknologi Sistem Informasi*, 6(1), 79–88. <https://doi.org/10.35957/jtsi.v6i1.9819>
- Anggraeni, R. F. D., Hasanah, N., Pahala, I., & Wahono, P. (2025). Kajian Terhadap Undang-Undang No 7 Tahun 2021 Tentang Harmonisasi Peraturan Perpajakan. *Jiip - Jurnal Ilmiah Ilmu Pendidikan*, 8(2), 1896–1900. <https://doi.org/10.54371/jiip.v8i2.7039>
- Annisawati, A. A., Kambali, I., & Yanto, R. T. Y. (2021). Analisis Time and Motion: Efisiensi Beban Kerja Direct Marketing Di Era Disrupsi Digital. *Eqien: Jurnal Ekonomi Dan Bisnis*, 8(2). <https://doi.org/10.34308/eqien.v8i2.268>
- Ardiningtyas, Y. E., Heruningsih, P., & Rosa, P. (2021). Analisis Balancing Data Untuk Meningkatkan Akurasi Dalam Klasifikasi. *Prosiding Seminar Nasional Aplikasi Sains & Teknologi (SNAST) 2021*, 2021(Prosiding SNAST 2021), A24–A28.
- Hartanto Dharmawan, A. (2023). Faktor penyebab pembatalan faktur pajak. *Jurnal Ilmiah Akuntansi Dan Keuangan*, 5(9), 2023. <https://journal.ikopin.ac.id/index.php/fairvalue>
- Iskandar, D., & Kurniawati, A. (2025). Analisis Perbandingan Teknik Word2vec dan Doc2vec dalam Mengukur Kemiripan Dokumen Menggunakan Cosine Similarity. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, 12(1), 133–144. <https://doi.org/10.25126/jtiik.2025129143>
- Khansyah, A. M., & Khotimah, S. (2025). the Role of Administration in the Process of Verification and Recapitulation of Ppn At Pt. Pelita Indonesia Djaya (Pelni Services). *Diploma International Conference 2025 FEB UNSOED*, 1–18.
- Kusnantoro, Rohana, T., & Sulistya Kusumaningrum, D. (2022). Implementasi Metode Tesseract OCR(Optical Character Recognition) untuk Deteksi Plat Nomor Kendaraan Pada Sistem Parkir. *Scientific Student Journal for Information, Technology and Science*, 3(1), 59–67.
- Meirza, A., & Puteri, N. R. (2024). Implementasi Metode YOLOV5 dan Tesseract OCR untuk Deteksi Plat Nomor Kendaraan. *Jurnal Ilmu Komputer Dan Desain Komunikasi Visual*, 9(1), 424–435.

- Muhammad Yahya, Wahyudi, & Akmal Hidayat. (2023). Implementasi Artificial Intelligence (AI) di Bidang Pendidikan Kejuruan Pada Era Revolusi Industri 4.0. *Seminar Nasional Dies Natalis 62, 1*, 190–199. <https://doi.org/10.59562/semnasdies.v1i1.794>
- Nashiruddin, M., Dhaza Noor Praditya, F., Faiz Mufazzal, A., & Ardiansyah, A. (2025). *Implementasi OCR berbasis Tesseract untuk Ekstraksi data kartu mahasiswa UMKLA. 01(1)*, 11–14.
- Nugraha, K. A. (2024). Penerapan Optical Character Recognition untuk Pengenalan Variasi Teks pada Media Presentasi Pembelajaran. *Jurnal Buana Informatika, 15(1)*, 69–78.
- Nurwicaksono, M. A., Lisa, I. N., Tiara, A. R., & Sidik, R. (2023). Optimasi Sistem Informasi Konsultasi Hukum melalui Pendekatan Pengujian Kombinasi White-box dan Black-box. *Jurnal Manajemen Informatika (JAMIKA), 14(1)*, 1–15. <https://doi.org/10.34010/jamika.v14i1.10110>
- Rahman, S., Sembiring, A., Siregar, D., Khair, H., Gusti Prahmana, I., Puspadini, R., & Zen, M. (2023). Python : Dasar Dan Pemrograman Berorientasi Objek. In *Penerbit Tahta Media*.
- Reyvansyah, M. R. (2023). Penerapan Metode Optical Character Recognition (OCR) Untuk Mengambil Data Arsip. *Jurnal Teknik Elektro Dan Komputer TRIAC, 10(2)*, 44–50. <https://doi.org/10.21107/triac.v10i2.20809>
- Rosnita, L., & Adek, R. T. (2023). Peningkatan Efisiensi Kecepatan dan Akurasi Rekapitulasi Faktur Pajak Dengan Optical Character Recognition Di Orbit Future Academy. *TECHSI - Jurnal Teknik Informatika, 14(2)*, 152. <https://doi.org/10.29103/techsi.v14i2.14861>
- Sukarna, R. H., Safira, F., & Marzuliyanti, H. D. (2025). *International Jurnal of Informatics Studi Komparatif Model OCR Berbasis AI untuk Dokumen Cetak dan Tulisan Tangan*. 1–11.
- Wahyuddin, W., & Hasim, A. (2023). Aplikasi Ekstraksi Data Kartu Vaksin Berbasis Web Menggunakan Metode Ocr. *Jurnal Sintaks Logika, 3(2)*, 53–57. <https://doi.org/10.31850/jsilog.v3i2.2525>