Use of Cosine Similarity, Manhattan Distance, and Jaccard Similarity Methods to Improve the Accuracy of Manual Payment Evidence Validation in ERP Applications

Sheilla Amira
Magister of Computer Science
Budi Luhur University
Jakarta, Indonesia
2311601757@student.budiluhur.ac.id

Muslim
Magister of Computer Science
Budi Luhur University
Jakarta, Indonesia
2311600734@student.budiluhur.ac.id

Wendi Usino
Magister of Computer Science
Budi Luhur University
Jakarta, Indonesia
wendi.usino@budiluhur.ac.id

Abstract— Manual validation of payment receipts in Enterprise Resource Planning (ERP) applications often faces challenges in terms of Accuracy, especially when payment data must be matched with existing transactions. Data mismatches can lead to recording errors and increase the burden of manual verification. This study aims to improve the Accuracy of payment receipt validation by comparing three Similarity methods: Cosine Similarity, Jaccard Similarity, and Manhattan Distance. In this research, Optical Character Recognition (OCR) is utilized to validate scanned images of payment receipts. By using OCR, data from receipt images can be automatically extracted into text format for further processing. The experimental results show that Cosine Similarity delivers the best performance, with a Precision of 100%, Recall of 90%, and Accuracy of 90%. On the other hand, Jaccard Similarity failed to identify any valid data, resulting in 0% across all evaluation metrics. Meanwhile, Manhattan Distance achieved high Precision (100%) but performed poorly in Recall and Accuracy, both at 10%. Based on these findings, Cosine Similarity is recommended as the most effective method for enhancing OCR-based payment validation in ERP systems. This study also opens the opportunity to develop hybrid approaches, combining Cosine Similarity and Manhattan Distance methods to further improve overall system performance.

Keywords— Payment Proof Validation, Cosine Similarity, Jaccard Similarity, Manhattan Distance, Accuracy

I. INTRODUCTION

Accurate financial transaction processing is critical for organizational efficiency, particularly within Enterprise Resource Planning (ERP) systems where manual validation of payment evidence remains a significant bottleneck. This process is prone to data mismatches, leading to recording errors, increased fraud risk, and substantial operational overhead [1]. While Optical Character Recognition (OCR) offers a viable means of digitizing payment receipts [2], and similarity matching algorithms are established tools for data comparison, their integration within ERP validation workflows lacks comprehensive analysis. Prior work often isolates these components—focusing either on OCR extraction [3] or applying a single similarity measure like Cosine Similarity in generic contexts [4]. This leaves a gap in understanding which similarity method performs optimally for the specific task of matching OCR-derived data to structured ERP transactions.

This study addresses that gap by conducting a direct comparative evaluation of three similarity algorithms—Cosine Similarity, Jaccard Similarity, and Manhattan Distance—within an integrated OCR-ERP validation framework. The novelty lies not in the individual methods, but in their empirical comparison for this specific application, guiding the development of targeted automation solutions. Cosine Similarity is selected for its effectiveness with text-based TF-IDF vectors [5], Jaccard Similarity for set-based comparisons [6], and Manhattan Distance for its robustness with numerical data like transaction amounts [7].

The primary objective is to benchmark these methods by evaluating their accuracy, precision, and recall in matching extracted payment data to the corresponding ERP transaction records. Performance will be assessed on a dataset of scanned payment receipts, with the goal of identifying the most reliable algorithm to reduce manual workload and enhance validation accuracy. By aligning methodological comparison with concrete ERP validation challenges and explicit evaluation metrics [8], this research aims to provide evidence-based guidance for implementing automated, accurate payment verification systems.

II. RESEARCH METHODOLOGY

To ensure the level of conformity between OCR extraction data and reference data stored in the ERP system, various similarity methods are used. These include Cosine Similarity, Jaccard Similarity, and Manhattan Distance. These three methods work with different approaches to measure the level of similarity between data. Cosine Similarity measures the angle between two vectors in multidimensional space to determine the similarity of direction or distribution of words/data [9]. Jaccard Similarity calculates the ratio of the intersection to the union of two sets, which is suitable for comparing data based on unique words or tokens. Manhattan Distance measures the total absolute difference between vector components, providing an indication of how much actual difference there is between two normalized numerical or text data.

A. Literature Review

By applying these three methods, the data validation system does not rely on a single approach, but can compare various aspects of similarity, including direction, token occurrence, and numerical distance. This improves the accuracy and reliability of the manual payment validation process in ERP applications [10]. The results of the literature review analysis are shown in Table 1 below.

Table 1. Literature Review Analysis

No	Authors and Years	Research Title	Method Used	Research Result
1	(Septio et al. 2023)	Development of a Digital Purchase Invoice Validation Application Using OCR	OCR technology and Tesseract tool	The application results were tested using Black Box testing, showing that all application functions worked well, thereby facilitating the digital billing validation process [11].
2	(Larsson and Segerås 2016)	Automated Invoice Handling with Machine Learning and OCR for Improved Data Validation	OCR technology and machine learning	The results show that the developed system can automatically detect and correct errors in invoice data with a high degree of accuracy, improving the company's operational efficiency [12].
3	(Azzam et al. 2023)	The Use of Blockchain Technology and OCR in E-Government for Document Management: Inbound Invoice Management as an Example	OCR technology and blockhain	The results of the study show that integrating OCR with blockchain can improve the security and accuracy of invoice management, as well as speed up the document validation process [3].
4	(SENTRIN 2020) dikutip dari Sesi Paralel : Sentrin 1 [218]	Analysis of the Cosine Similarity Method in Automatic Online Essay Examination Applications (Case Study of JTI Polinema)	Cosine Similarity	The results of the study show that the application of this method improves the accuracy of automatic essay scoring on online exam platforms. The use of Cosine Similarity allows for a more accurate comparison between participants' answers and the expected answers. Accuracy testing using Precision,

				Recall, and f-measure resulted in an average accuracy of 81% [13] .
5	(SENTRIN 2020) dikutip dari Sesi Paralel : Sentrin 8 [212]	Comparison of Convolutional Neural Network-Based Skin Lesion Prescreening: Original and Segmented Images	Convolutional Neural Network (CNN)	The results show that image segmentation improves the accuracy of CNN models in detecting skin lesions, with validation accuracy increasing from 0.82 to 0.90 despite using smaller training data (22.41% of the total data). This study also proves that prescreening skin lesions with CNN does not require image background removal [13]
6	(SENTRIN 2020) dikutip dari Sesi Paralel : Sentrin 1 [283]	Sentiment Analysis of Coffee Shop Reviews Using the Naïve Bayes Method with Genetic Algorithm Feature Selection	Naïve Bayes and Genetic Algorithm for Features Selection	The results of the evaluation of Naïve Bayes classification with feature selection using the Genetic Algorithm show an accuracy of 0.944, precision of 0.945, recall of 0.944, and f-measure of 0.945, with the best parameters being 50 generations, 18 populations, a crossover rate of 1, and a mutation rate of 0. This study shows that feature selection with a genetic algorithm significantly improves sentiment prediction accuracy compared to without feature selection [13].
7	(Halim and Lasut 2024)	Document Plagiarism Detection Application Using Web-Based TF-IDF and Cosine Similarity Methods (English)	Combination of TF-IDF and Cosine Similarity	The result of the journal article is an application that uses the Cosine Similarity and TF-IDF methods to process the similarity values of the documents tested. The use of stemming in the calculation will produce higher similarity values. The Cosine Similarity method is widely used in text similarity calculations due to its high level of accuracy [14].
8	(Wang, Chen, and Hu 2022)	Multi-View Cosine Similarity Learning with Application to Face Verification	Multi-View (e.g HOG, LBP, CNN) and Cosine Similarity	The result of this article is that the proposed method provides higher accuracy compared to other face verification methods, especially in challenging conditions such as variations in lighting, pose, and facial expressions [15].
9	(Baruah et al. 2023)	Exploring Jaccard Similarity and Cosine Similarity for Developing an Assamese Question-Answering System	Data collection, data pre-processing (tokenization, stopword removal, stemming), Jaccard similarity, Cosine similarity.	The results of comparing the Cosine Similarity and Jaccard Similarity methods are as follows: Cosine Similarity, supplemented with text pre-processing, achieved the highest correlation with a total score of 93%, while Jaccard Similarity achieved a score of 87% [6]
10	(Adi 2024)	Implementation of Handwriting Recognition Using Optical Character Recognition with CNN and RNN Methods on Receipts and Invoices	OCR CNN and RNN method	The results of this study using a combination of CNN and RNN models achieved an accuracy of 83.33% in recognizing nominal numbers on handwritten receipts, demonstrating the potential of deep learning methods in OCR systems. The combination of

				CNN and RNN is effective in recognizing handwriting on receipts and invoices, although accuracy still needs to be improved [11].

From previous studies, the author applied Optical Character Recognition (OCR) technology and compared three Similarity algorithms, namely Cosine Similarity, Jaccard Similarity, and Manhattan Distance in this study because all three have significant potential to improve accuracy and efficiency in the manual payment proof validation process. OCR technology is used to automatically extract data from payment receipt images [16], while the three Similarity methods are applied to match the extraction results with reference data in the ERP system, using different similarity calculation approaches.

The novelty of this study lies in the comparison of the performance of the three Similarity methods in the context of ERP systems, which has not been systematically explored in previous studies. Additionally, this research uses a realistic dataset consisting of various formats and qualities of payment evidence, thus approximating real-world conditions. It is hoped that the results of this research can provide practical contributions to the automation of data validation processes in ERP systems, reducing manual workload, and improving the accuracy and reliability of financial transaction recording.

B. Research Object Scope and Context

This research focuses on the validation of payment evidence in the context of Enterprise Resource Planning (ERP) applications. Payment evidence validation is an important process in financial management, which ensures that transactions recorded in the system are accurate and consistent with supporting documents. In this study, the methods used to improve accuracy are Cosine Similarity, Jaccard Similarity, and Manhattan Distance, which are techniques in data processing to measure the similarity between two or more documents [17].

This research was conducted in the context of companies that have implemented ERP systems but still experience problems with the accuracy of manual payment validation. By integrating OCR and using the Similarity method (cosine, jaccard, and manhattan), it is hoped that errors can be reduced and accuracy in the validation process can be improved.

C. Framework Concept

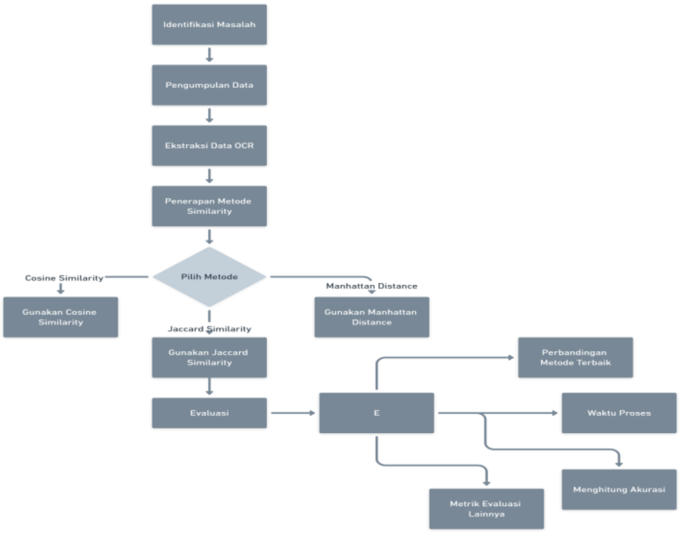


Figure 1. Framework Concept

1. Problem Identification

Analyze challenges in the manual payment proof validation process and data from the ERP system.

2. Data Collection
   Collect manual payment proof data in the form of images and actual transaction data from the ERP system.
3. OCR Data Extraction
   Extract text from image-based payment proofs into numerical and text data [18].
4. Similarity Method Application
   Develop and implement Cosine Similarity, Jaccard Similarity, and Manhattan Distance algorithms to compare payment evidence with ERP data.
5. Evaluate Results
   a. Calculate accuracy using each Similarity method.
   b. Perform evaluation using evaluation metrics such as Precision, Recall, and Accuracy.
   c. Perform comparative analysis to determine the best method.
   d. Calculating the processing time for each method.

D. Cosine Similarity

Here are the steps in calculating Cosine Similarity.

1. Two texts to be compared. For example, “12345678” and “12345679”.
2. Preprocessing, the text cleaning stage, such as removing punctuation marks, removing stop words, and stemming.
3. Vectorization: Converting the processed text into numerical vectors using methods such as TF-IDF or Word2Vec. Examples of vector results:
   - OCR Vector: [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]
   - Reference Vector: [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]
4. Cosine Similarity Calculation, Calculating the Cosine Similarity value between two vectors representing text.
5. Similarity Score, The result of the Cosine Similarity calculation, which is a value between 0 and 1. Example: Cosine Similarity accuracy result: 0.97 - Clear, Match
6. Threshold, Determines whether the score indicates a high or low level of similarity, based on a predetermined threshold.
   - 1.00: Very Clear, Match. Identical or nearly identical text, with no character differences
   - 0.95 – 0.99: Clear, Match. Very similar, only minor differences (spacing, capital letters)
   - 0.90 – 0.94 : Fairly Clear, Needs to be Checked Again. Similar but there are small differences in words/phrases.
   - 0.70 – 0.89 : Less Clear, Needs to be Checked Again. Quite significant differences, could be incorrect names/accounts
   - < 0.70 : Not Clear, Not a Match

E. Jaccard Similarity

Jaccard Similarity is a method for measuring how similar two sets are by calculating the ratio between the number of elements in the intersection and the union. In the context of payment evidence validation, this method is particularly suitable for comparing token-shaped entities such as account numbers or recipient bank names that consist of character fragments. For example, if two strings differ due to minor OCR errors such as ‘12345678’ and ‘12345679’, Jaccard Similarity will still recognize the similarity of most of the digits. This method is considered quite effective in handling cases where the extracted data is not completely identical but has many similar elements, such as a single digit error in an account number. The Jaccard Similarity formula is as follows.

$$J(A,B) = (A \cap B)/|A \cup B| \quad (1)$$

Where:

- $|A \cap B|$ is the number of elements that appear in both sets A and B.
- $|A \cup B|$ is the total number of unique elements in set A or B or both.

The Jaccard value ranges from 0 (no similarity) to 1 (identical). The closer it is to 1, the greater the proportion of elements that are the same in the two sets.

F. Manhattan Distance

Manhattan Distance, also known as City Block Distance, is a method of measuring the distance between two points by summing the absolute differences of each dimension. In this study, this method is very relevant for comparing numerical values such as payment amounts between OCR results and data in the ERP system. For example, if the OCR result reads Rp. 243,908 while the ERP system records Rp. 243,900, Manhattan Distance will give a difference value of 8. The main advantages of Manhattan Distance are its simplicity and sensitivity to numerical differences, making it ideal for detecting small differences that may not be captured by text-based methods. This method does not take into account direction or vectors, but only the total absolute difference in relevant dimensions.
The Manhattan Distance formula is as follows:

$$J(A,B) = \Sigma\, |Ai - Bi| \quad (2)$$

Where:

- Ai and Bi are elements of two vectors A and B
- Σ is the total sum of the absolute differences between the corresponding components

Research by (Paulo and Paulo 2024) proves that this approach is very effective for numerical data validation in digital transactions, as it provides intuitive and easy-to-interpret measurements of small errors.

III. DISCUSSION

This research uses an experimental quantitative approach, aiming to compare the performance of three methods: Cosine Similarity, Jaccard Similarity, and Manhattan Distance in the process of validating manual payment evidence in an ERP system. This approach is used to test the accuracy of the system in matching OCR-extracted data with actual transaction data, through the calculation of evaluation values such as Precision, Recall, and Accuracy. This research was conducted with the following steps:

1. Data Collection, Collecting a dataset of manual payment evidence from the ERP application.
2. OCR Implementation, Using OCR to extract text from images.
3. Similarity Validation, Implementing the Cosine Similarity, Jaccard Similarity, and Manhattan Distance methods in payment evidence validation.
4. Performance Evaluation,
    a. Measuring the performance of the three Similarity methods using Precision, Recall, and Accuracy metrics to determine their effectiveness in validating payment evidence.

b. Analyzing the effectiveness of the three methods based on the evaluation results in the metrics to determine the most optimal method for validating manual payment evidence.

In this study, three Similarity methods were used to compare their effectiveness in the manual payment evidence validation process in the ERP application. The Cosine Similarity method was chosen because it is capable of measuring similarities between texts by considering the angles between vectors in multidimensional space, making it effective for matching OCR extraction results from payment receipt images with textual transaction reference data, such as recipient names and account numbers [19].

Jaccard Similarity was used to measure similarity based on token or character sets, which allows for measuring the level of overlap between OCR results and original data, especially for short texts that are prone to character recognition errors. Meanwhile, Manhattan Distance is used to calculate the total absolute difference between numerical values, and is particularly relevant when comparing numerical features such as payment amounts, taking into account small error tolerances due to OCR noise. With a comparative approach to these three methods, the study aims to identify the most optimal method for improving the accuracy of payment proof validation in ERP systems, especially when handling OCR extraction data that contains both text and numerical information [20].

A. Sample Selection Method

The data used in this study consists of manual payment evidence that has been processed and validated through the ERP system. This payment evidence can be in the form of transaction receipts or bank transfer evidence sent by customers, and is stored digitally (in .jpg or .png format) in the company's ERP system. This data was used as the main object in the process of evaluating and comparing the effectiveness of three Similarity methods, namely Cosine Similarity, Manhattan Distance, and Jaccard Similarity, in the context of text and numeric-based payment validation.

Sample selection was carried out using purposive sampling techniques, which is selecting samples based on certain criteria that have been determined in advance. The selected payment evidence consists of documents with sufficient image quality to be processed using Optical Character Recognition (OCR) technology. In addition, samples are also selected based on the completeness of important information required for the data matching process, such as the recipient's name, account number, and payment amount. These three elements are important to ensure that the results extracted from the images can be accurately compared with the transaction data stored in the ERP system [21].

The inclusion criteria in this study included payment evidence in the form of clear images with text that could be easily recognized by OCR, as well as containing complete information relevant to the ERP system. Meanwhile, the exclusion criteria included payment evidence that was blurred or unreadable, did not contain valid transaction information, or had been modified in such a way that it was not possible to carry out a valid data validation or comparison process [22].

The amount of data used in this study was 30 payment proof samples that met the inclusion criteria. The selection of 30 payment evidence samples was based on the experimental approach used in this study. The main focus of the study was on comparing the performance of the Cosine Similarity, Manhattan Distance, and Jaccard Similarity algorithms in the context of OCR-based payment evidence validation. The number of 30 samples was considered representative for the initial experiment, as it covered the variations in form, nominal value, and format of payment evidence commonly found in ERP systems. Furthermore, this approach has been used in various previous studies that focused on evaluating machine learning methods or similarity metrics with limited data but in-depth analysis. Thus, even though the amount of data is not large, this study still makes a significant contribution to testing the accuracy and effectiveness of validation methods in real-world scenarios [23].

B. Data Collection

The data used in this study will be collected from two sources:

1. Primary data: data will be collected from payment receipts uploaded by app users, which may be in the form of transaction receipts or bank transfer receipts in image format (.jpg or .png). These documents will be the main objects extracted using OCR (Optical Character Recognition) technology to obtain text information such as the recipient's name, account number, and payment amount.

2. Secondary data: purchase transaction data in the ERP system, where the data is contained on a page that contains information on the payment amount corresponding to the purchase invoice. This data is used as a reference or comparative data in the validation process, which will be matched with the results of data extraction from proof of payment using the Cosine Similarity, Manhattan Distance, and Jaccard Similarity methods.

Both types of data are needed to support the process of comparing the effectiveness of the three Similarity methods in improving the accuracy of manual proof of payment validation, in accordance with the main focus of this research.

C. Analysis Technique

Data extracted from payment receipt images using OCR technology (Amazon Textract) will be analyzed using three similarity methods, namely Cosine Similarity, Jaccard Similarity, and Manhattan Distance. The goal is to determine the level of similarity between the OCR-extracted text data and the transaction reference data contained in the ERP system. For example, if the extraction results are represented as vector A and the reference data as vector B, the calculation process is carried out using an approach appropriate to each method.

In the Cosine Similarity method, the text will be converted into a vector representation to measure the angle of similarity between vectors, while Jaccard Similarity compares the intersection and union of characters/words from two strings to determine the percentage of similarity. Manhattan Distance is used to calculate the total absolute difference between numerical features such as nominal and account numbers.

After the Similarity calculation is performed and the document is classified as valid or invalid, the next step is to evaluate the performance of each method using a Confusion Matrix. Through the Confusion Matrix, evaluation metrics such as Precision, Recall, and Accuracy will be calculated to determine which method has the best performance in the manual payment proof validation process in the ERP application [24]. The matrix calculation is explained in the following figure.

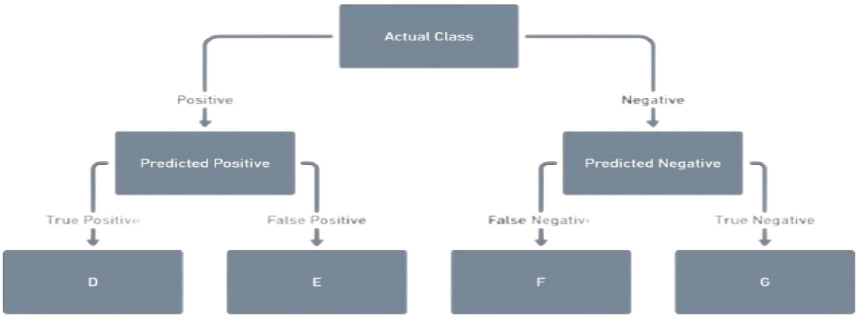


Figure 2. Confusion Matrix Flowchart

The following is an explanation of the Confusion Matrix.

- TP (True Positive): relevant results that are correctly predicted as relevant.
- FP (False Positive): irrelevant results that are incorrectly predicted as relevant.
- FN (False Negative): relevant results that are incorrectly predicted as irrelevant.
- TN (True Negative): irrelevant results that are correctly predicted as irrelevant.

From Figure 1 above, Precision, Recall, and Accuracy can be calculated using the following formulas.

1. Precision

Measures the proportion of valid results that are correct out of all validation results that are considered correct.

$$\frac{True\ Positives\ (TP)}{True\ Positives\ (TP) + False\ Positives\ (FP)} \tag{1}$$

- True Positive (TP): Correct prediction for the positive class.
- False Positive (FP): Incorrect prediction for the positive class (actually negative).

2. Recall

Measures the proportion of truly valid documents that are successfully recognized.

$$\frac{True\ Positive\ (TP)}{True\ Positive\ (TP) + False\ Negative\ (FN)} \tag{2}$$

- False Negative (FN): Positive cases that are incorrectly predicted as negative.

3. Accuracy

Measuring the percentage of correct validation results from the overall data.

$$\frac{True\ Positives\ (TP) + True\ Negatives\ (TN)}{Total\ PrPr\ e\ diction} \tag{3}$$

D. Data Design and Testing

The following stages of the design are visualized in the following image.

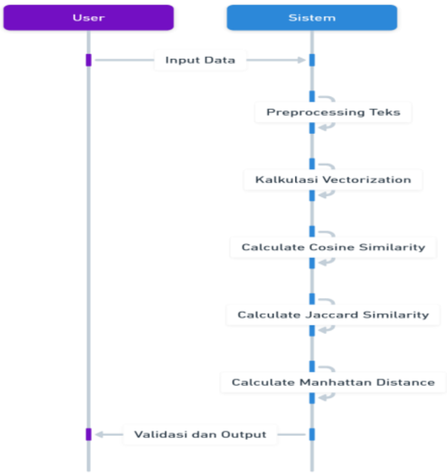

User
Sistem
Input Data
Preprocessing Teks
Kalkulasi Vectorization
Calculate Cosine Similarity
Calculate Jaccard Similarity
Calculate Manhattan Distance
Validasi dan Output

Figure 3. Data Testing Stage

The following is an explanation of the manual payment receipt validation data testing stage in an ERP application that uses a comparison approach of three methods: Cosine Similarity, Jaccard Similarity, and Manhattan Distance.

1. Data Input

   Users upload payment receipts in image format (.jpg/.jpeg) to the ERP system. The image contains information such as the recipient's name, account number, and payment amount.

2. OCR Process

   The system extracts text from the image using OCR (Amazon Textract) technology. The extraction results are then compared with the reference data stored in the ERP database. The steps are as follows:

   - Text Preprocessing

     The OCR and reference texts will undergo preprocessing, such as:

     - Removing punctuation marks and excess spaces
     - Text tokenization, example of tokenization results:

[“m-transfer”, “successful”, “02/25/2025”, “16:08:00”, “to”, “3730082222”, “leu”, “retail”, “indonesia”, “pt”, ‘rp’, “34240”]

- Vectorization Calculation
  The processed data is then converted into vector form for similarity calculation.

3. Cosine Similarity
   Convert text into vectors based on term frequency or TF-IDF [13], then calculate the similarity angle between vectors.
   $$\text{Cosine Similarity} = (A \bullet B) / (\|A\| * \|B\|) = 12 / (\sqrt{12} * \sqrt{12}) = 1.0 = 100\% \text{ similarity}$$
   If there are slight differences, such as errors in numbers or spaces, the similarity value can drop to 0.98 or 98% similarity.
4. Jaccard Similarity
   Calculates the ratio between the intersection and union of tokens.
   $$\text{Jaccard} = |A \cap B| / |A \cup B|$$
5. Manhattan Distance
   Used to compare numerical features.
   $$\text{Manhattan Distance} = \Sigma |xi - yi|$$
6. Validation and Output
   The results of the three Similarity methods are used to validate whether the payment evidence matches the reference data in the ERP. The system will display:
   - Scores from Cosine, Jaccard, and Manhattan
   - Validation conclusion: valid, needs re-checking, invalid
   - Accuracy status: Very Clear, Clear, Fairly Clear, Less Clear, Not Clear

System testing was conducted in three stages, namely:

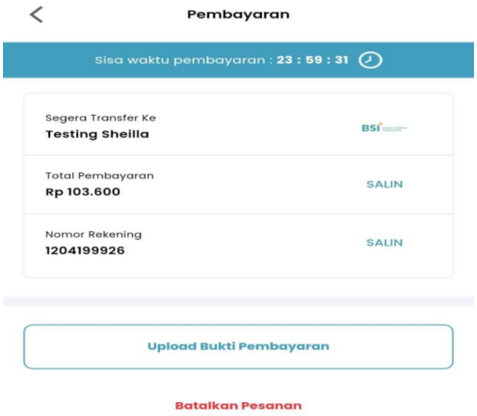
1. OCR Accuracy Testing
   Evaluation based on image extraction results and comparison with the original text.
2. Similarity Method Testing
   Calculating Cosine Similarity, Jaccard Similarity, and Manhattan distance values and analyzing the performance of each method in determining the suitability of payment evidence.
3. Validation System Testing
   Measuring the processing time for each method, starting from the OCR stage – preprocessing – Similarity calculation – to validation output.

This study used 30 samples of payment evidence images in .jpg format uploaded through the mobile ERP application. The inclusion criteria included images that were clear and readable by OCR. The samples represented manual transactions carried out by users. The validation process was carried out to ensure that the information from the images matched the references in the ERP system. The data used in this study was then evaluated using three different similarity methods: Cosine Similarity, Jaccard Similarity, and Manhattan Distance to obtain an objective comparison of validation performance. The processed data had the following characteristics.

- Format: .jpg and .jpeg images
- Test Data: Recipient name, account number, and amount
- Quality criteria: Images must have sufficient resolution for OCR to extract text accurately

E. Validation Process

ERP users upload proof of payment images through a feature in the mobile application. The images are sent to the ERP central server for automatic validation. The upload feature is available in the form of an Upload Proof button on the mobile application, where users can select images from the gallery or take photos directly using the camera.



Pembayaran

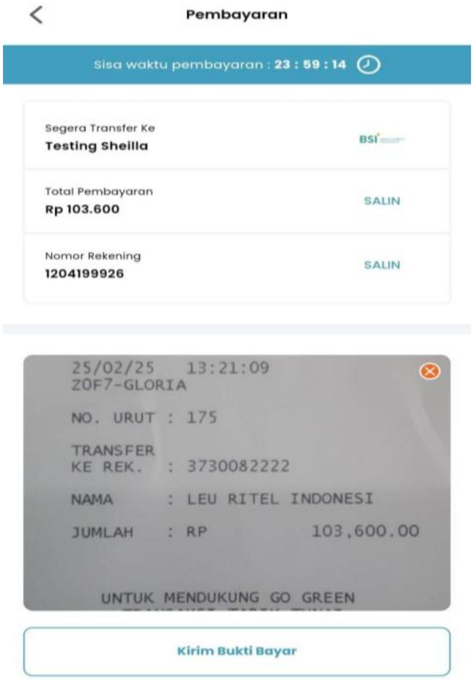
Sisa waktu pembayaran : 23 : 59 : 31

Segera Transfer Ke
Testing Sheilla

Total Pembayaran
Rp 103.600

Nomor Rekening
1204199926

Figure 4. Upload proof of payment page display



Pembayaran

Sisa waktu pembayaran : 23 : 59 : 14

Segera Transfer Ke
Testing Sheilla

Total Pembayaran
Rp 103.600

Nomor Rekening
1204199926

```
25/02/25    13:21:09
ZOF7-GLORIA

NO. URUT : 175

TRANSFER
KE REK.    : 3730082222

NAMA       : LEU RITEL INDONESI

JUMLAH     : RP          103,600.00


UNTUK MENDUKUNG GO GREEN
```

Figure 5. Preview page display

After the image is received by the ERP system, the validation process is carried out automatically on the server side. The process includes:

1. Text Extraction (OCR)
   The ERP system runs an OCR engine on image files uploaded from the mobile application.
2. Text Preprocessing
   Cleaning of OCR text results to remove noise and foreign characters, and tokenization.
3. Similarity Check
   Calculating the similarity of OCR text with the reference using three methods: Cosine Similarity, Jaccard Similarity, Manhattan Distance.
4. Validation Assessment

The system determines whether the proof of payment is Valid or Invalid based on the threshold value for each Similarity method.

5. Output to ERP Web UI (Admin)
   The similarity results will be displayed in the report. The system calculates the three methods in parallel to support the evaluation and comparison of each algorithm's performance.

IV. RESEARCH RESULT

The results of the validation process can be viewed by the ERP administrator via the dashboard page for all data or in reports for transaction data results. The text extraction process is carried out using Optical Character Recognition (OCR) technology on 30 payment receipt images in .jpg format uploaded by application users and obtained through the ERP system. The purpose of this process is to convert the information in the images into text that can be further analyzed by comparing methods in the validation process using the Cosine Similarity, Jaccard Similarity, and Manhattan Distance algorithms.

After the accuracy results are obtained, the average accuracy for each data component will be calculated. The following table shows the OCR extraction accuracy results from a total of 30 test data.

Table 2. OCR Accuracy Result

Component	Accuracy Average	Total Accurate	Total Not Accurate
Recipient Name	73.33%	22	8
Account Number	100%	30	0
Pay Amount	100%	30	0

Based on the evaluation results of 90 data components from 30 payment receipts (each consisting of the Recipient Name, Account Number, and Payment Amount), the OCR system showed excellent performance in recognizing text from payment receipt images. The Account Number and Payment Amount components were successfully recognized with 100% accuracy across all 30 samples. This shows that numeric characters with a consistent format tend to be more easily recognized by the OCR system without distortion or reading errors.

Meanwhile, for the Recipient Name component, the system was able to read accurately in 22 out of 30 samples (73.33%). Some discrepancies were found due to truncated company names. Overall, the OCR system was able to accurately extract data from the majority of important elements of the payment receipt. The high accuracy rate, especially for the Account Number and Payment Amount, demonstrates the system's reliability in handling crucial information in the payment validation process. These results provide a strong foundation for further validation using the Cosine Similarity, Jaccard, or Manhattan Distance algorithms to test accuracy against ERP reference data.

The following picture shows the results of the Cosine Similarity calculation on 30 samples.

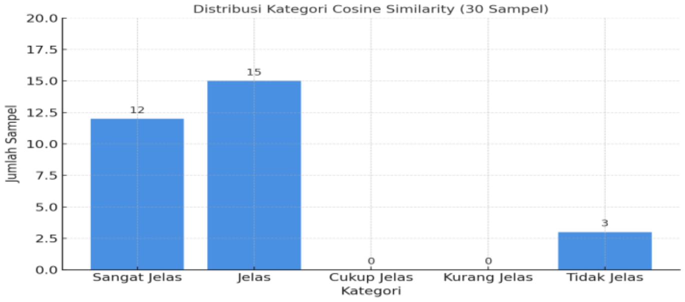


Figure 6. Cosine Similarity Distribution Validation

The following picture shows the results of the Jaccard Similarity calculation on 30 samples.

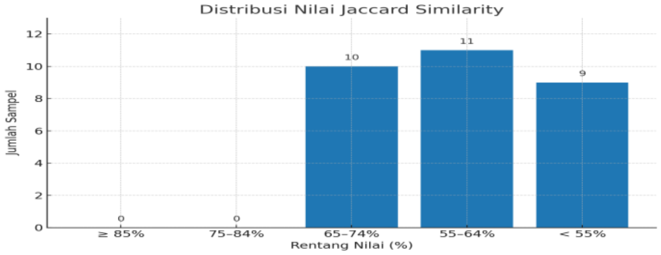


Figure 7. Jaccard Similarity Distribution Validation

The following picture shows the results of the Manhattan Distance calculation on 30 samples.

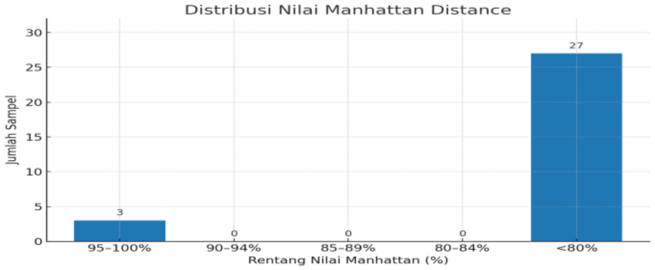


Figure 8. Manhattan Distance Distribution Validation

To evaluate the effectiveness of each method in validating payment evidence, a comparison was made between the following methods.

- Cosine Similarity
- Jaccard Similarity
- Manhattan Distance

Performance evaluation focuses not only on validation accuracy, but also on the speed and efficiency of the system in processing payment evidence. This is important because it will have a direct impact on user experience, especially in ERP systems that are actively used by many users. The system was tested using 30 payment evidence samples. Measurements were taken from the moment the payment evidence was uploaded to the system until the validation result (valid/invalid) was obtained.

Table 3. Performance results per method

ID Transaction	Component Method	OCR Process (seconds)	*Similarity Check* (seconds)	Total Process (seconds)
TRX-001	Cosine	1.57	0.01	2.02
TRX-001	Jaccard	1.35	0.02	1.81
TRX-001	Manhattan	1.40	0.03	1.86
TRX-002	Cosine	1.78	0.04	2.54
TRX-002	Jaccard	1.56	0.03	2.05
TRX-002	Manhattan	1.58	0.04	2.07
TRX-003	Cosine	1.72	0.04	1.79
TRX-003	Jaccard	1.65	0.02	1.65
TRX-003	Manhattan	1.74	0.05	1.74
TRX-004	Cosine	1.56	0.02	2.15
TRX-004	Jaccard	1.59	0.03	2.10
TRX-004	Manhattan	2.07	0.05	2.55
TRX-005	Cosine	2.21	0.04	2.86
TRX-005	Jaccard	2.32	0.02	3.11

TRX-005	Manhattan	2.13	0.03	2.71
TRX-006	Cosine	1.63	0.03	1.63
TRX-006	Jaccard	1.53	0.04	1.54
TRX-006	Manhattan	1.47	0.06	1.47
TRX-007	Cosine	1.65	0.03	1.73
TRX-007	Jaccard	1.50	0.04	1.50
TRX-007	Manhattan	1.52	0.05	1.52
TRX-008	Cosine	1.80	0.04	1.88
TRX-008	Jaccard	1.72	0.03	1.72
TRX-008	Manhattan	1.81	0.04	1.81
TRX-009	Cosine	1.53	0.04	1.60
TRX-009	Jaccard	1.60	0.04	1.60
TRX-009	Manhattan	2.02	0.03	2.02
TRX-010	Cosine	1.33	0.01	1.33
TRX-010	Jaccard	1.38	0.02	1.38
TRX-010	Manhattan	1.39	0.06	1.39
TRX-011	Cosine	1.73	0.04	1.80
TRX-011	Jaccard	1.77	0.04	1.77
TRX-011	Manhattan	1.72	0.05	1.72
TRX-012	Cosine	1.48	0.02	1.48
TRX-012	Jaccard	1.33	0.04	1.33
TRX-012	Manhattan	1.49	0.06	1.49
TRX-013	Cosine	1.71	0.02	2.34
TRX-013	Jaccard	1.61	0.02	2.13
TRX-013	Manhattan	1.56	0.03	2.07
TRX-014	Cosine	2.99	0.03	3.59
TRX-014	Jaccard	2.24	0.03	2.24
TRX-014	Manhattan	2.34	0.08	2.34
TRX-015	Cosine	1.84	0.03	1.84
TRX-015	Jaccard	1.94	0.02	1.94
TRX-015	Manhattan	2.98	0.05	2.98
TRX-016	Cosine	2.65	0.05	2.99
TRX-016	Jaccard	2.54	0.03	2.54
TRX-016	Manhattan	2.78	0.06	2.78
TRX-017	Cosine	2.54	0.04	2.71
TRX-017	Jaccard	2.57	0.04	2.57
TRX-017	Manhattan	2.39	0.06	2.39
TRX-018	Cosine	2.14	0.06	2.26
TRX-018	Jaccard	1.93	0.06	1.93
TRX-018	Manhattan	2.03	0.07	2.03
TRX-019	Cosine	2.55	0.04	4.00
TRX-019	Jaccard	2.63	0.04	3.65
TRX-019	Manhattan	2.72	0.10	4.19
TRX-020	Cosine	1.73	0.01	1.73

TRX-020	Jaccard	1.74	0.03	1.74
TRX-020	Manhattan	1.63	0.04	1.63
TRX-021	Cosine	4.10	0.03	5.81
TRX-021	Jaccard	3.11	0.02	4.10
TRX-021	Manhattan	3.53	0.03	5.08
TRX-022	Cosine	1.68	0.01	2.27
TRX-022	Jaccard	1.83	0.02	2.47
TRX-022	Manhattan	1.89	0.05	2.78
TRX-023	Cosine	1.65	0.02	1.65
TRX-023	Jaccard	2.41	0.03	2.41
TRX-023	Manhattan	2.48	0.04	2.48
TRX-024	Cosine	2.05	0.02	3.22
TRX-024	Jaccard	1.98	0.05	2.63
TRX-024	Manhattan	3.15	0.05	3.98
TRX-025	Cosine	1.82	0.02	1.82
TRX-025	Jaccard	2.67	0.02	2.67
TRX-025	Manhattan	1.89	0.09	1.89
TRX-026	Cosine	2.94	0.03	4.67
TRX-026	Jaccard	2.67	0.03	4.46
TRX-026	Manhattan	3.12	0.04	4.41
TRX-027	Cosine	3.94	0.03	4.19
TRX-027	Jaccard	2.89	0.04	2.90
TRX-027	Manhattan	2.59	0.04	2.59
TRX-028	Cosine	3.26	0.04	4.08
TRX-028	Jaccard	2.66	0.05	2.66
TRX-028	Manhattan	2.78	0.07	2.78
TRX-029	Cosine	2.89	0.03	5.99
TRX-029	Jaccard	3.28	0.05	4.78
TRX-029	Manhattan	2.94	0.06	4.00
TRX-030	Cosine	3.17	0.03	3.62
TRX-030	Jaccard	2.75	0.03	2.75
TRX-030	Manhattan	2.62	0.09	2.62

Table 4. Process time average

Method	OCR Process Average (seconds)	*Similarity* Process Average (seconds)	Total Average Process (seconds)
Cosine	2.18	0.03	2.72
Jaccard	2.09	0.03	2.40
Manhattan	2.19	0.05	2.51

Cosine Similarity has the longest total processing time (2.72 seconds) even though its Similarity time is relatively fast (0.03 seconds). Jaccard Similarity is the most efficient overall with the lowest total processing time (2.40 seconds). Manhattan Distance has the slowest Similarity processing (0.05 seconds), but its total processing time is still below Cosine.

V. Conclusion

Based on the results of the research conducted, summarized in the following results:

1. The Cosine Similarity method is the most effective and balanced method, both in terms of accuracy and processing speed. This method produces Precision of 100%, Recall of 90%, and Accuracy of 90%, which demonstrates the system's ability to accurately and consistently recognize valid data. Although the average total processing time is slightly longer than other methods (2.72 seconds), its high accuracy performance makes this method highly recommended for primary implementation.
2. The verification of the conformity between the nominal value of the OCR extraction results and the transaction value in the ERP was successful, with a high similarity value (e.g., Cosine $\geq$1.00 or Manhattan having no nominal difference) proving that the system is capable of accurately matching numerical data automatically.
3. The main challenges in OCR-based data matching lie in format variations, noise, and the possibility of character or number reading errors. However, the use of similarity methods with appropriate text preprocessing can reduce the impact of these errors.
4. Of the three methods compared, Cosine Similarity is the most consistent in handling text-based data, while Manhattan Distance is more sensitive in comparing numerical data. Jaccard similarity shows low performance on data with few differences, especially if the tokenization elements are not optimal.
5. The addition of a literature review on sentiment analysis strengthens the theoretical basis of the research, as the concepts of text processing, tokenization, and similarity measurement are also important parts of sentiment analysis. This shows that a text-based approach in the NLP domain can be effectively applied to ERP validation systems.

Acknowledgement

We would like to express our deepest gratitude to the Computer Sains of Budi Luhur University lecturer, friends and family for their support and trust to support this research. And for those who involved in helping us complete this research.

References

[1] Arief, M., & Rafik, A. (2024). Pengelolaan Proyek Implementasi ERP pada Sistem Laporan Keuangan Parkir di PT. Harfan Tri Megah (Edugate). 02(06), 79–92.

[2] Alhadian, F. (2024). Analisis Perencanaan Penerapan Enterprise Resource Planning (ERP) dalam Aktivitas Manajerial di Yayasan Sosial dan Pendidikan Bina Muda Cicalengka. 2(2).

[3] Azzam, F., Jaber, M., & Saies, A. (2023). applied sciences The Use of Blockchain Technology and OCR in E-Government for Document Management: Inbound Invoice Management as an Example.

[4] Jiang, P. (2024). A Survey of Text-Matching Techniques.

[5] Representation, F., & Similarity, C. (2024). Journal of Dinda. 4(2), 149–153.

[6] Baruah, N., Gupta, S., Ghosh, S., & Afrid, S. N. (n.d.). Exploring Jaccard Similarity and Cosine Similarity for Developing an Assamese Question Answering System. 1–11.

[7] Paulo, S., & Paulo, S. (n.d.). Asymptotic behavior of the Manhattan distance in ? -dimensions : Estimating multidimensional scenarios in empirical experiments.

[8] Shade, B., & Altmann, E. G. (2023). Quantifying the Dissimilarity of Texts.

[9] Rahmawati, V., Julianty, L., Fauziah, S., & Rafif, S. (2025). A Comparative Study of Cosine Similarity and Manhattan Distance on Text Representations Using TF-IDF

and Bag of Words. 210–215. https://doi.org/10.1109/ICITCOM66635.2025.11265201

[10] Amelia, N., Agama, I., Negeri, I., & Kerinci, I. (2024). Eksplorasi Validitas dan Reliabilitas Soal Pemahaman Konsep dalam Asesmen Pembelajaran. 2(1), 222–232.

[11] Komputer, J. S., Septio, P. A., Yulianto, S., Prasetyo, J., Kristen, U., & Wacana, S. (2023). Pembuatan Aplikasi Validasi Document Tagihan Pembelian Barang Secara Digital Menggunakan OCR dengan tool tesseract pada System Portal Perusahaan. 7(September), 650–662.

[12] Larsson, A. (2016). Automated invoice handling with machine learning and OCR Automatiserad fakturahantering med maskininlärning och OCR.

[13] Timur, J., Komputer, F. I., & Brawijaya, U. (n.d.). Sentrin 2020.

[14] Halim, J., & Lasut, D. (2024). Document Plagiarism Detection Application Using Web-Based TF-IDF and Cosine Similarity Methods. 7(2). https://doi.org/10.32877/bt.v7i2.1697

[15] Wang, Z., Chen, J., & Hu, J. (2022). Multi-View Cosine Similarity Learning with Application to Face Verification. 1–13.

[16] Adi, C. (2024). Implementasi Pengenal Tulisan Tangan Menggunakan Optical Character Recognition Dengan Metode Cnn Dan Rnn Pada Dokumen Resi Dan Kuitansi. 11(1), 32–38.

[17] Amer, A. A., & Abdalla, H. I. (2020). A set theory based similarity measure for text clustering and classification. Journal of Big Data. https://doi.org/10.1186/s40537-020-00344-3

[18] Fauziah, S., Saputra, D. D., Pratiwi, R. L., & Kusumayudha, M. R. (2023). Komparasi Metode Feature Selection Text Mining Pada Permasalahan Klasifikasi Keluhan Pelanggan Industri Telekomunikasi Menggunakan Smote Dan Naïve Bayes. IJIS - Indonesian Journal On Information System, 8(2), 174. https://doi.org/10.36549/ijis.v8i2.289

[19] Agustina, T., Masrizal, M., & Irmayanti, I. (2024). Performance Analysis of Random Forest Algorithm for Network Anomaly Detection using Feature Selection. Sinkron, 8(2), 1116–1124. https://doi.org/10.33395/sinkron.v8i2.13625

[20] Ronzon, T., Gurria, P., Carus, M., Cingiz, K., El-Meligi, A., Hark, N., Iost, S., M'barek, R., Philippidis, G., van Leeuwen, M., Wesseler, J., Medina-Lozano, I., Grimplet, J., Díaz, A., Tejedor-Calvo, E., Marco, P., Fischer, M., Creydt, M., Sánchez-Hernández, E., ⋯ Miras Ávalos, J. M. (2025). No 主観的健康感を中心とした在宅高齢者における 健康関連指標に関する共分散構造分析Title. Sustainability (Switzerland), 11(1), 1‐14. https://doi.org/10.1016/j.resenv.2025.100208%0A

[21] Wang, Z., Cai, Z., & Wu, Y. (2023). An improved YOLOX approach for low-light and small object detection: PPE on tunnel construction sites. Journal of Computational Design and Engineering, 10(3), 1158–1175. https://doi.org/10.1093/jcde/qwad042

[22] Yu, L., Yang, X., Wei, H., Liu, J., & Li, B. (2024). Driver fatigue detection using PPG signal, facial features, head postures with an LSTM model. Heliyon, 10(21). https://doi.org/10.1016/j.heliyon.2024.e39479

[23] Dharmendra, I. K., Agus, I. M., Putra, W., & Atmojo, Y. P. (2024). Evaluasi Efektivitas SMOTE dan Random Under Sampling pada Klasifikasi Emosi Tweet. Informatics for Educators And Professionals : Journal of Informatics, 9(2), 192–193.

[24] Jupin, J. A., Sutikno, T., Ismail, M. A., Mohamad, M. S., Kasim, S., & Stiawan, D. (2019). Review of the machine learning methods in the classification of phishing attack. Bulletin of Electrical Engineering and Informatics, 8(4), 1545–1555. https://doi.org/10.11591/eei.v8i4.1344