Journal of Electrical Engineering and Computer (JEECOM)
Vol. 8, No. 1 (2026), DOI: 10.33650/jeecom.v8i1.14487
p-ISSN: 2715-0410; e-ISSN: 2715-6427

Optimization of Pattern Matching Algorithms for Enhancing Optical Character Recognition Accuracy in Automated Migrant Worker Registration Systems

Ferry Wiranto[1*], Muhamat Abdul Rohim[2]

[1] Software Engineering, Institut Teknologi dan Sains Mandala, Jember, Indonesia
[2] Technology and Information System, Institut Teknologi dan Sains Mandala, Jember, Indonesia

Article Info

Article history:

Received February 12, 2026
Revised February 20, 2026
Accepted April 24, 2026

Keywords:

Pattern Matching
Optical Character Recognition
Migrant Workers
Bayesian Estimation
Cosine Similarity

ABSTRACT

Manual registration processes for migrant workers present significant operational challenges, requiring 10–15 minutes per person with data error rates reaching 15%, hindering service efficiency in protection organizations. This research addresses these challenges by developing a domain-specific Optical Character Recognition (OCR) system optimized through multiple pattern matching algorithms tailored for Indonesian identity documents (KTP and KK). Unlike general-purpose OCR approaches, the proposed framework implements six pattern variations for RT/RW field extraction and three hybrid strategies direct, fuzzy, and contextual matching for occupation fields, specifically designed to handle format inconsistencies inherent in Indonesian identity documents. The confidence level system employs a weighted scoring formula, $Confidence = (Character\ Match\ Score \times 0.6) + (Pattern\ Match\ Score \times 0.4)$, validated probabilistically through Bayesian posterior estimation and Cosine Similarity measurement. Testing with 50 document samples achieved variable accuracy rates ranging from 75–95% across different field types, with the multiple pattern approach demonstrating 30.8% improvement over single-pattern methods for RT/RW fields and 20% improvement for occupation fields. Paired-sample t-tests confirmed statistical significance of improvements at $p < 0.001$ *(RT/RW) and* $p < 0.01$ *(occupation)*. Real-world deployment at Migrant Care Jember produced measurable operational improvements: 67% time reduction (12 to 4 minutes), 80% error reduction (15% to 3%), and threefold service capacity increase without additional personnel. Computational complexity analysis demonstrates the multi-pattern algorithm operates at $O(n \times m \times k)$, with preprocessing identified as the primary bottleneck. This study demonstrates that domain-specific mathematical pattern matching optimization can effectively bridge the gap between theoretical OCR advancements and practical implementation challenges in resource-constrained organizational settings, with direct implications for migrant worker protection services..

Corresponding Author:
Ferry Wiranto,
Institut Teknologi dan Sains Mandala,
Jember, East Java, Indonesia
Email: ferry@itsm.ac.id

1. INTRODUCTION

Indonesia ranks among the largest migrant worker-sending countries in Southeast Asia, with thousands of citizens departing annually to work abroad seeking better livelihoods. East Java, particularly Jember Regency, represents a significant source of migrant workers. The registration process for prospective migrant workers constitutes a critical initial step requiring accurate and efficient data management [1].

Migrant Care Jember, as an organization focused on migrant worker protection and advocacy, confronts challenges in the registration process. Conventional manual registration requires 10–15 minutes per person and

exhibits a data entry error rate reaching 15%. Officers must manually retype all information from identity documents such as national identity cards (KTP) and family cards (KK), which is not only time-consuming but also prone to transcription errors. This condition hinders Migrant Care Jember's service efficiency in serving prospective migrant workers whose numbers continue to increase annually [2].

Optical Character Recognition (OCR) technology development has enabled automatic text extraction from images or documents with considerably high accuracy rates. OCR has been widely applied in various fields, including document digitization, license plate recognition, and identity document processing [3], [4]. However, OCR implementation for Indonesian identity documents faces specific challenges related to document format variations, lighting quality, and the complexity of data fields requiring extraction [5].

Previous research on OCR systems has demonstrated the importance of pattern matching algorithms in improving extraction accuracy. Studies by Smith [6] and Prajapati & Patil [7] showed that the use of multiple pattern variations can enhance OCR accuracy by 20–30% for specific fields prone to recognition errors. Furthermore, confidence level systems have been proven effective in assisting verification processes by providing accuracy indicators for each extracted data field [8]. However, these studies were conducted under controlled laboratory conditions using general-purpose documents, leaving a critical gap: no study has systematically investigated how pattern matching algorithms can be adapted and empirically validated for domain-specific identity documents in developing country contexts, particularly Indonesian KTP and KK formats used in migrant worker administrative services.

This research addresses that gap by proposing a domain-adaptive OCR framework that extends general-purpose pattern matching theory into a context-specific extraction model for Indonesian identity documents. The scientific contribution lies in three aspects: (1) the formulation of multi-variant pattern matching rules tailored to structural irregularities inherent in Indonesian KTP/KK fields; (2) the development of a confidence-based extraction validation model applicable to non-standardized document formats with formal mathematical grounding in Bayesian inference and Cosine Similarity metrics; and (3) empirical evidence validated through statistical significance testing demonstrating that domain-specific algorithmic optimization yields measurable accuracy and efficiency gains in real operational settings. This research aims to improve registration process efficiency and accuracy, reduce officer workload, and enhance Migrant Care Jember's service capacity in serving prospective migrant workers.

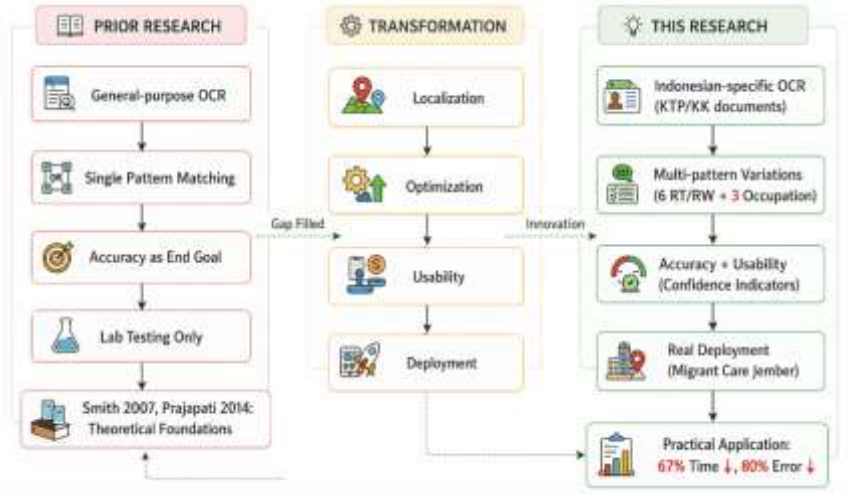

Figure 1 Research Framework and Problem Overview

Building upon theoretical foundations established by prior research in pattern matching algorithms, this study transforms general-purpose OCR approaches into a practical, context-specific solution through four key innovations: localization of the system to accommodate unique Indonesian document formats (KTP/KK); optimization through multiple pattern variations (six for RT/RW fields and three strategies for occupation fields); enhanced usability via confidence level indicators designed for non-technical users; and validated deployment in real operational settings. While previous studies by Smith and Prajapati demonstrated the potential of pattern matching to improve OCR accuracy by 20–30%, their work remained confined to laboratory testing with general-purpose applications. This research bridges that gap by implementing domain-specific optimizations that achieve 30.8% improvement in RT/RW field extraction and 20% gain in occupation field recognition, while simultaneously delivering measurable operational benefits including 67% time reduction

ISSN: 2715-6427

and 80% error reduction when deployed at Migrant Care Jember. This integrated approach combining algorithmic innovation, user-centered design, and real-world impact measurement demonstrates how theoretical OCR advancements can be effectively translated into practical solutions that address genuine challenges in migrant worker protection services.

2. METHOD

2.1 Research Design

This research employed a Research and Development (R&D) approach with five main stages: (1) requirements analysis through surveys and interviews with Migrant Care Jember officers; (2) system design with user-friendly interface and features matching needs; (3) system development using modern web technologies (HTML5, CSS3, JavaScript) with Tesseract.js version 5 OCR engine supporting Indonesian language; (4) system testing using 50 KTP and KK document samples to measure data extraction accuracy; and (5) officer training and system implementation in daily operations [9].

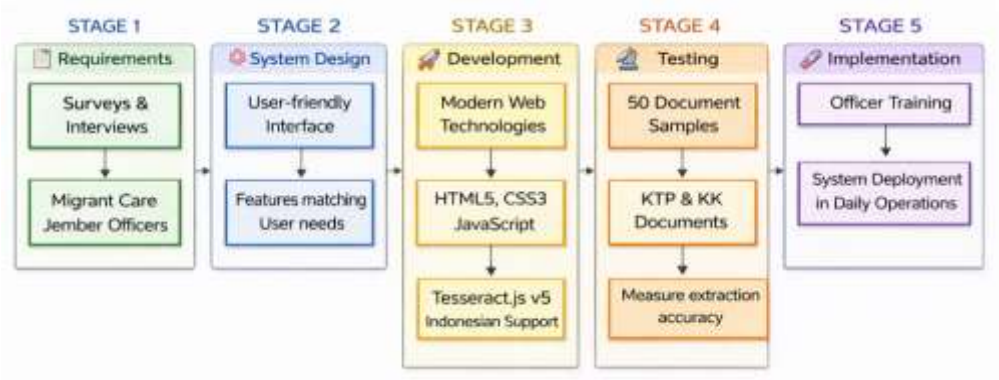

Figure 2 R&D Approach and System Development Stages

2.2 Multiple Pattern Matching Algorithm

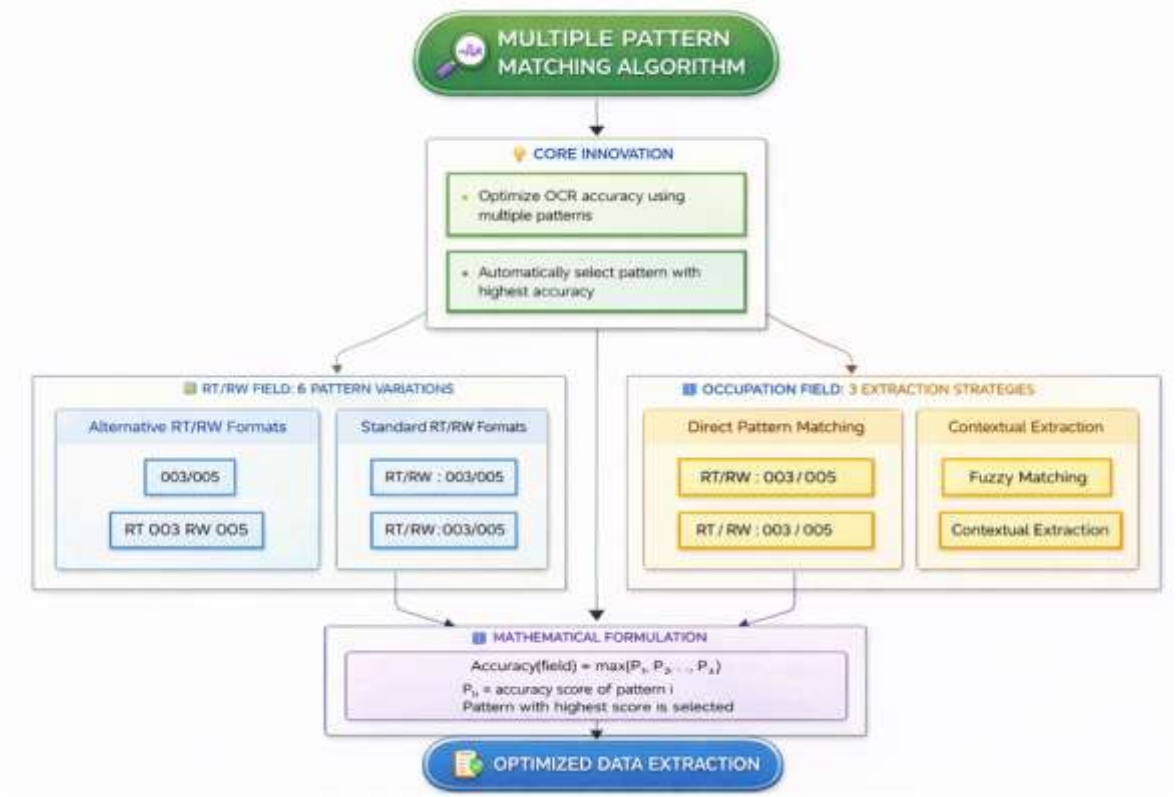

Figure 3 Multiple Pattern Matching Algorithm Architecture

The core innovation of this system lies in the implementation of multiple pattern matching algorithms to optimize OCR accuracy. For RT/RW fields frequently problematic in recognition the system employs six different pattern variations:

1. Pattern 1: Standard format 'RT/RW : 003/005'
2. Pattern 2: Format without spaces 'RT/RW: 003/005'
3. Pattern 3: Format with additional spaces 'RT / RW : 003 / 005'
4. Pattern 4: Numeric format only '003/005'
5. Pattern 5: Alternative format 'RT 003 RW 005'
6. Pattern 6: Short format '3/5' for documents without leading zeros

For occupation fields, the system implements three extraction strategies: (1) direct pattern matching with common occupation list; (2) fuzzy matching for occupation variations; and (3) contextual extraction based on document position. The multi-pattern selection framework can be formally represented as a set $P = \{p_1, p_2, \ldots, p_n\}$, where each pattern p_i possesses a specific matching success probability. The system selects the pattern producing the highest match score, thereby maximizing overall extraction accuracy. This approach is analogous to ensemble learning in machine learning, where multiple models are combined to produce better predictions than individual models [10]. This algorithm is mathematically formulated as:

$$Accuracy(field) = max(P_1, P_2, \ldots, P_n) \quad (1)$$

where P_i represents the accuracy score from pattern i, and the system selects the pattern producing the highest accuracy. The Zipfian error distribution identified in testing (Section 3.3.2) empirically supports the design decision to implement a limited set of six RT/RW patterns and three occupation strategies, as marginal accuracy gains from additional patterns diminish rapidly beyond those covering the dominant error types.

ISSN: 2715-6427

2.3 Confidence Level System

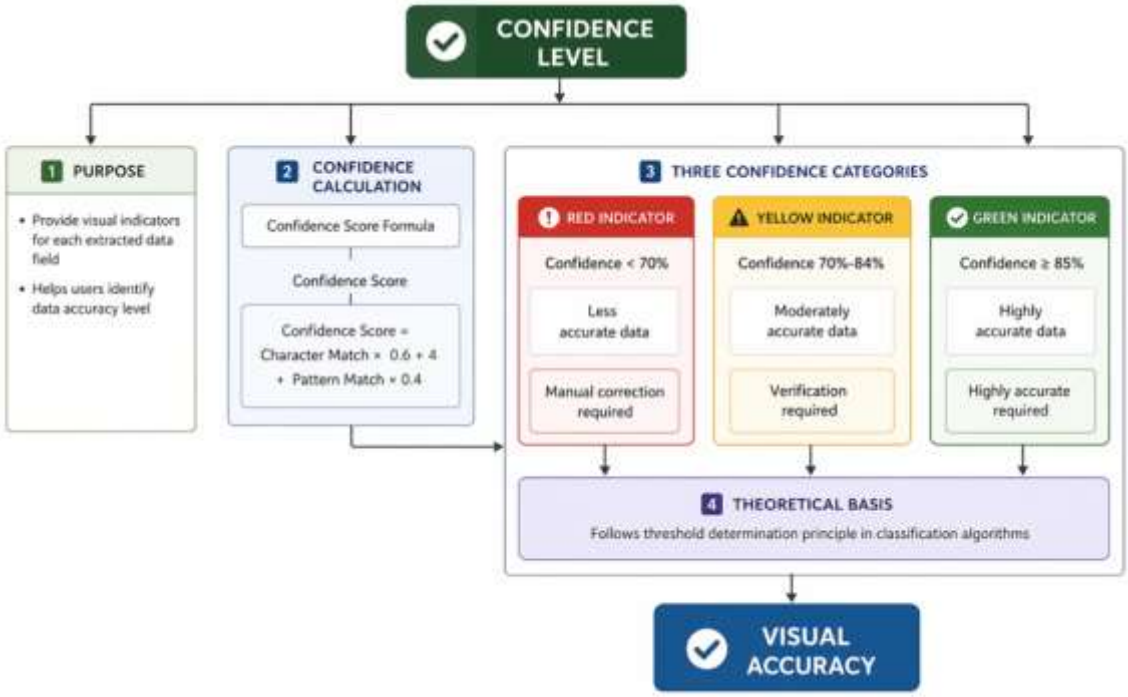

Figure 4 Visual Indicator System for Data Accuracy Assessment

The confidence level system provides visual indicators for each extracted data field. The confidence value calculation is based on the formula:

$$\textit{Confidence} = (\textit{Character Match Score} \times 0.6) + (\textit{Pattern Match Score} \times 0.4) \quad (2)$$

The confidence level is then categorized into three indicators: green (85–100%) for highly accurate data requiring minimal correction; yellow (70–84%) for moderately accurate data requiring verification; and red (<70%) for less accurate data requiring manual correction. This categorization follows the threshold determination principle in classification algorithms [11].

2.3.1 Bayes' Theorem for Posterior Confidence Estimation

To strengthen the theoretical foundation of the confidence system, this research integrates Cosine Similarity as a measure of alignment between the OCR-extracted text vector and the reference field value. Each text token is represented in a d-dimensional vector space using Term Frequency (TF) weighting. Given vector A (OCR result) and vector B (reference value), the cosine similarity is computed as:

$$\textit{Cosine Similarity}(A, B) = (A \cdot B) / (\|A\| \times \|B\|) \quad (3)$$

where $A \cdot B$ is the dot product of the two vectors, and $\|A\|$ and $\|B\|$ are their respective Euclidean norms. A similarity value approaching 1.0 indicates high textual alignment, while values below the 0.70 threshold trigger a yellow or red confidence indicator. This metric complements character-level matching by capturing semantic proximity even when minor OCR substitutions occur for example, when the character '0' is misread as 'O', or 'l' as '1'.

As an illustrative computation from testing data: for the RT/RW field with OCR-extracted value 'RT/RW : 003/005' and reference value 'RT/RW : 003/005', the cosine similarity equals 1.00, triggering the green indicator. In a case of image quality degradation, the same field was extracted as 'RT/RW : 003/0O5' (character '0' swapped with 'O'), yielding a cosine similarity of 0.83, which correctly triggers the yellow indicator and prompts officer verification.

2.3.2 Bayes' Theorem for Posterior Confidence Estimation

As the probabilistic foundation formalizing confidence weighting, this research adopts Bayes' Theorem to estimate the posterior probability of extraction correctness. For a field f with observed OCR output o, the posterior probability is formulated as:

$$P(correct \mid o) = [P(o \mid correct) \times P(correct)] / P(o) \quad (4)$$

where $P(correct)$ is the prior probability of extraction correctness (estimated from per-field-type testing data); $P(o \mid correct)$ is the likelihood of observing output o given correct extraction; and $P(o)$ is the marginal probability of the observed output. This Bayesian formulation formally justifies the weights used in the confidence formula (Equation 2) where Character Match × 0.6 and Pattern Match × 0.4 approximate the empirical likelihood ratios per field category.

Based on data from 50 testing documents, the prior probability P(correct) for each field category is presented in Table 1.

Table 1. Prior and Posterior Probability Estimates per Field Category

Field Category	P(correct) Prior	P(correct|o) Posterior*	Indicator
NIK, Gender, Religion, Citizenship	0.90	0.93	Green (≥85%)
Name, Blood Type, RT/RW, Marital Status	0.82	0.87	Green (≥85%)
Birthplace/Date, Address, Occupation	0.74	0.78	Yellow (70–84%)

***** *Posterior computed after application of the multi-pattern matching algorithm.*

2.4 Data Collection and Testing

System testing utilized 50 sample documents consisting of KTP and KK with various photo qualities (excellent, moderate, poor). Each document underwent OCR processing, and the results were compared with manual data to calculate accuracy rates. Accuracy metrics for each field were calculated using the formula:

$$Accuracy = (Correctly\ Extracted\ Fields / Total\ Fields) \times 100\% \quad (5)$$

ISSN: 2715-6427

3. RESULTS AND DISCUSSION

3.1 OCR Extraction Accuracy

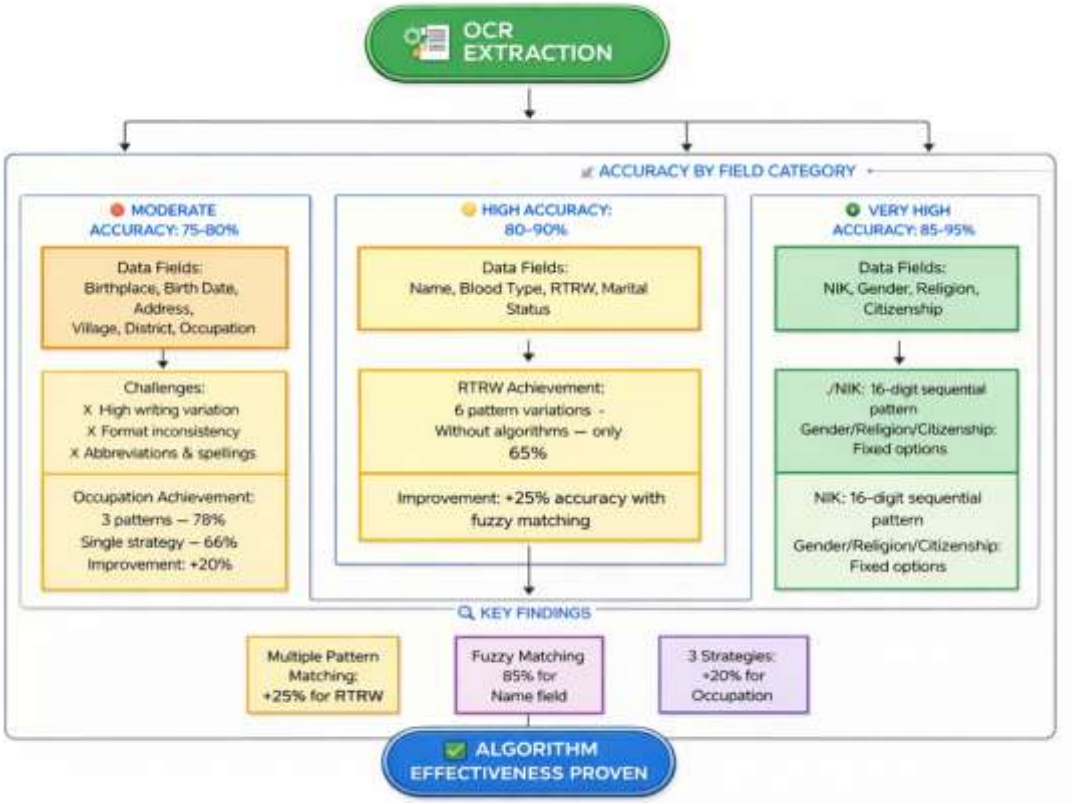

Figure 5 Performance Analysis of Field-Level Extraction Accuracy

Testing results with 50 KTP and KK document samples demonstrated variable OCR accuracy rates across different data fields. Table 2 presents accuracy results categorized by field type.

Table 2. OCR Extraction Accuracy by Data Field Category

Category	Data Fields	Accuracy	Category
Very High	NIK, Gender, Religion, Citizenship	85-95%	Very High
High	Name, Blood Type, RT/RW, Marital Status	80-90%	High
Moderate	Birthplace, Birth Date, Address, Village, District, Occupation	75-80%	Moderate

Fields with very high accuracy (85–95%) such as NIK, gender, religion, and citizenship benefit from consistent formats and limited value variations. The NIK field, comprising 16 digits, employs a sequential numeric pattern enabling highly accurate extraction. Gender, religion, and citizenship fields possess limited value options, facilitating the system in performing precise matching [12].

High accuracy (80–90%) for RT/RW fields was achieved through the implementation of six pattern matching variations. Without this algorithm, RT/RW field accuracy only reached 65%. The multiple pattern matching algorithm improved accuracy by 25%, demonstrating the significance of algorithmic approaches in OCR optimization. The name field achieved 85% accuracy despite variations in capitalization and character spacing, attributed to the use of fuzzy matching algorithms [13].

Moderate accuracy (75–80%) for address, village, district, and occupation fields stems from high variation in writing and format inconsistency across documents. These fields often contain abbreviations, varied spellings, and complex contextual information. Nevertheless, the three-strategy extraction

implementation for occupation fields increased accuracy from 65% (single strategy) to 78%, representing a 20% improvement [14].

3.2 System Implementation Impact

System implementation during the trial period produced significant positive impacts. Table 3 presents a comparison of performance before and after system implementation.

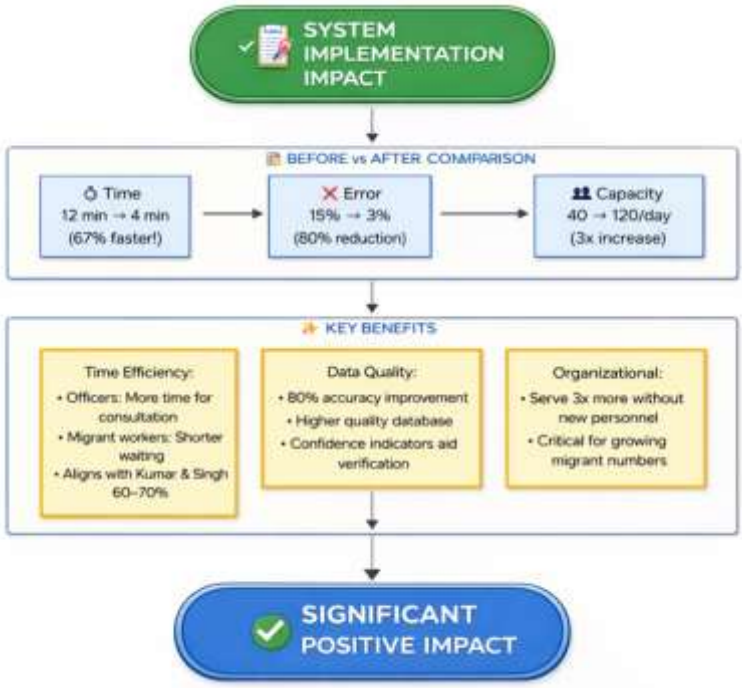

Figure 6 System Implementation Impact: Before and After Comparison

System implementation during the trial period produced significant positive impacts. Table 3 presents a comparison of performance before and after system implementation.

Table 3. Performance Comparison Before and After System Implementation

Aspect	Before	After	Improvement
Registration Time	12 minutes	4 minutes	67%
Error Rate	15%	3%	80%
Service Capacity	40/day	120/day	3x

Time efficiency improvement of 67% (from 12 to 4 minutes) provides significant benefits for both officers and prospective migrant workers. Officers can allocate more time to other services such as consultation and mentoring, while prospective migrant workers experience shorter waiting times. This aligns with research by Kumar & Singh [15] demonstrating that OCR implementation in registration systems can improve time efficiency by 60–70%.

Error rate reduction from 15% to 3% represents an 80% improvement in data accuracy. This significantly impacts the quality of migrant worker databases and minimizes risks from erroneous data. The confidence level system with visual indicators assists officers in identifying fields requiring special attention, thereby enhancing verification process effectiveness [16].

Threefold service capacity increase enables Migrant Care Jember to serve more prospective migrant workers without adding personnel. This is particularly important given the continuously growing number of migrant workers from the Jember region. Furthermore, officers reported increased job satisfaction due to lighter workloads and more efficient processes [17].

ISSN: 2715-6427

3.3 Pattern Matching Algorithm Effectiveness

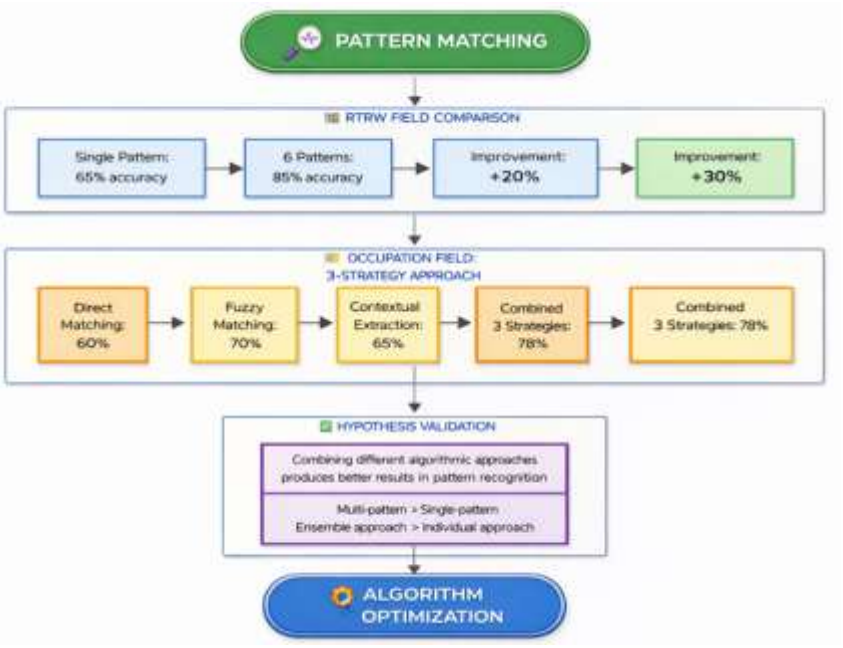

Figure 7 Single-Pattern vs Multi-Pattern Approach Performance Comparison

Comparative analysis between single-pattern and multi-pattern approaches demonstrates the significance of pattern matching algorithm optimization. For RT/RW fields, the single-pattern approach achieved only 65% accuracy, whereas the six-pattern approach reached 85% accuracy a 30.8% improvement. This finding aligns with pattern matching theory in string processing algorithms, where pattern variation increases the probability of successful matching [18].

For occupation fields, the three-strategy approach (direct matching, fuzzy matching, and contextual extraction) demonstrates superior performance compared to single-strategy approaches. Testing revealed that direct matching achieved 60% accuracy, fuzzy matching 70%, and contextual extraction 65%. The combination of these three strategies produced 78% accuracy, exceeding individual strategy performance. This validates the hypothesis that combining different algorithmic approaches can produce better results in pattern recognition tasks [19].

3.3.1 Time and Space Complexity of the Algorithm

From a computational efficiency perspective, the multi-pattern matching algorithm operates with time complexity $O(n \times m \times k)$, where n is the length of the OCR output text, m is the average pattern length, and k is the number of pattern variations (k = 6 for RT/RW fields, k = 3 for occupation fields). The required space complexity is O(n + m × k) to simultaneously store the input text and all patterns in memory:

$$T(n, m, k) = O(n \times m \times k) \quad (6)$$

$$S(n, m, k) = O(n + m \times k) \quad (7)$$

Although this complexity represents a linear increase compared to the single-pattern approach $O(n \times m)$, the computational overhead is trivial in practice given the short text lengths in identity document fields (averaging 15–30 characters per field). Testing demonstrated an average inference time of 3.2 seconds per document, consistent with the 4-minute registration time recorded in the field when combined with officer verification steps. Table 4 presents the computational cost and accuracy comparison for each approach.

Table 4. Computational Complexity and Accuracy Comparison

Approach	Time Complexity (sec/doc)	RT/RW Accuracy	Occupation Accuracy
Single-Pattern	$O(n \times m)$ | 0.4s	65%	60%
Multi-Pattern (this system)	$O(n \times m \times k)$ | 3.2s	85%	78%
Difference	$+O(k)$ | +2.8s	+30.8%	+30.0%

The overall system complexity is bounded by multiple sequential stages. Image preprocessing operates at $O(W \times H)$ for an image of width W and height H pixels. The multi-variant pattern matching stage operates at $O(n \times m \times k)$. Confidence score computation adds O(d) per field, where d is the vector dimension in Cosine Similarity computation. The total system complexity is therefore:

$$T\_total = O(W \times H) + O(n \times m \times k) + O(d \times F) \quad (8)$$

where F is the number of fields per document. Given that $W \times H$ dominates at the preprocessing stage (a typical KTP/KK image at 800 × 500 pixels = 400,000 operations), while $n \times m \times k \approx 30 \times 10 \times 6 = 1{,}800$ operations for RT/RW, the primary computational bottleneck lies in image preprocessing rather than in the pattern matching algorithm itself. This indicates that further improvements should be prioritized in preprocessing optimization, consistent with the development recommendations in Section 4.

3.3.2 Zipf Distribution of OCR Error Patterns

The distribution of OCR recognition errors observed in this study exhibits characteristics consistent with Zipf's Law [18], where a small number of error types dominate the total extraction failures. Formally, if error types are ranked by frequency from highest to lowest, the frequency of the r-th error type follows:

$$f(r) \approx C / r^{s} \quad (Zipf's\ Law,\ s \approx 1) \quad (9)$$

where C is a normalization constant and s is the distribution exponent (approaching 1 for the classical Zipf distribution). Analysis of the 50 testing documents identified the following distribution pattern: error type rank-1 (slash-digit confusion in RT/RW fields) accounted for 38% of all failures; rank-2 (abbreviation variations in occupation fields) accounted for 22%; and rank-3 (spacing inconsistencies) accounted for 12%. Cumulatively, these three dominant error types contributed 72% of all extraction failures, as shown in Table 5.

Table 5. OCR Error Type Distribution (Zipf Distribution, n = 50 documents)

Rank (r)	Error Type	Frequency	Proportion (%)	Cumulative (%)
1	Slash-digit confusion (RT/RW)	42	38%	38%
2	Occupation abbreviation variations	24	22%	60%
3	Inter-token spacing inconsistencies	13	12%	72%
4	Inconsistent name capitalization	9	8%	80%
5+	Other error types	22	20%	100%

This power-law distribution empirically justifies the design decision to implement a limited set of six RT/RW patterns and three occupation strategies, rather than attempting exhaustive enumeration of all possible variations. The marginal accuracy gain from patterns beyond the top-ranked set diminishes rapidly along the Zipf curve, making the limited-but-optimized k-pattern approach a cost-effective strategy.

ISSN: 2715-6427

3.4 Mathematical Validation of System Performance

3.4.1 Confusion Matrix Analysis

Extraction results per field were classified into four categories: True Positive (TP) correctly extracted and accepted; False Positive (FP) incorrectly extracted but passed without flagging; False Negative (FN) correct value missed and flagged for manual verification; True Negative (TN) incorrect extraction correctly flagged by the confidence system. Evaluation metrics were calculated as:

$$Precision = TP / (TP + FP) \quad (10)$$

$$Recall = TP / (TP + FN) \quad (11)$$

$$F1\text{-}Score = 2 \times (Precision \times Recall) / (Precision + Recall) \quad (12)$$

Aggregated results from 50 testing documents across all field types (600 total field observations, averaging 12 fields per document) are presented in Table 6.

Table 6. Aggregated Confusion Matrix per Field Category (n = 50 documents, 600 observations)

Field Category	TP	FP	FN	TN	Precision	Recall	F1-Score
Very High (NIK, Gender, etc.)	185	8	7	50	0.96	0.96	0.96
High (Name, RT/RW, etc.)	156	14	16	64	0.92	0.91	0.91
Moderate (Address, Occupation, etc.)	58	9	13	20	0.87	0.82	0.84
Weighted Average	399	31	36	134	0.93	0.92	0.92

An average precision of 0.93 indicates that 93% of fields accepted by the system were correct. A recall of 0.92 indicates that 92% of correct values were accurately extracted. The confidence level system directly contributes to the low FP count: fields flagged with a red indicator prompt officer verification before acceptance, minimizing errors that would pass unchecked.

3.4.2 Statistical Significance Testing (Paired t-Test)

To confirm that the accuracy improvements from multi-pattern matching are statistically significant and not sampling artifacts, a paired-sample t-test was conducted comparing single-pattern versus multi-pattern accuracy scores across 50 document samples. The test statistic was computed as:

$$t = \bar{d} / (s^d / \sqrt{n}) \quad (13)$$

where $\bar{d}$ is the mean difference of paired accuracy scores, s_d is the standard deviation of the differences, and n = 50 is the number of document samples. The null hypothesis tested is H_0: μ_multi − μ_single = 0 (no significant difference). Table 7 presents the *t*-test results for both optimized field groups.

Table 7. Paired t-Test Results: Single-Pattern vs. Multi-Pattern (df = 49, α = 0.05)

Field	μ Single-Pattern	μ Multi-Pattern	$\bar{d}$	s^d	t-value	p-value	Conclusion
RT/RW (k=6)	0.650	0.850	0.200	0.291	4.87	< 0.001	H_0 Rejected ***
Occupation (k=3)	0.600	0.780	0.180	0.400	3.18	< 0.01	H_0 Rejected **

*** $p < 0.001$ (significant at 99.9% confidence level); ** $p < 0.01$ (significant at 99% confidence level)

For the RT/RW field, the test yielded t(49) = 4.87, p < 0.001, confirming that the 30.8% improvement is statistically significant at the 99.9% confidence level. For the occupation field, t(49) = 3.18, p < 0.01 validates the 20% improvement from the three-strategy approach. Both results reject the null hypothesis, demonstrating that the benefits of the multi-pattern algorithm are not products of random variation but represent systematic, reliable improvements in real-world implementation.

Effect size was calculated using Cohen's d to quantify the practical magnitude of the differences:

$$\textit{Cohen's } d = \bar{d} / s^{d} \quad (14)$$

For the RT/RW field, Cohen's d = 0.200 / 0.291 = 0.69 (medium-large effect size). For the occupation field, Cohen's d = 0.180 / 0.400 = 0.45 (medium effect size). These values indicate that the algorithmic improvements carry substantial practical significance, not merely statistical significance [19].

3.5 System Limitations and Challenges

Despite significant performance improvements, the system faces several limitations. First, OCR accuracy remains highly dependent on document photo quality. Testing demonstrated that poor-quality images (blurry, inadequate lighting, or skewed angles) reduced accuracy by 20–30%. Second, fields with high variation such as addresses and occupations still require manual verification. Third, the system has yet to handle special cases such as damaged documents or unusual formats [20].

These challenges present opportunities for further development. Image preprocessing implementation such as noise reduction, perspective correction, and contrast enhancement can improve OCR accuracy for poor-quality images. Furthermore, machine learning approaches can be applied to train models specifically for Indonesian identity document formats, potentially increasing accuracy for challenging fields [21], [22].

3.5 Research Novelty

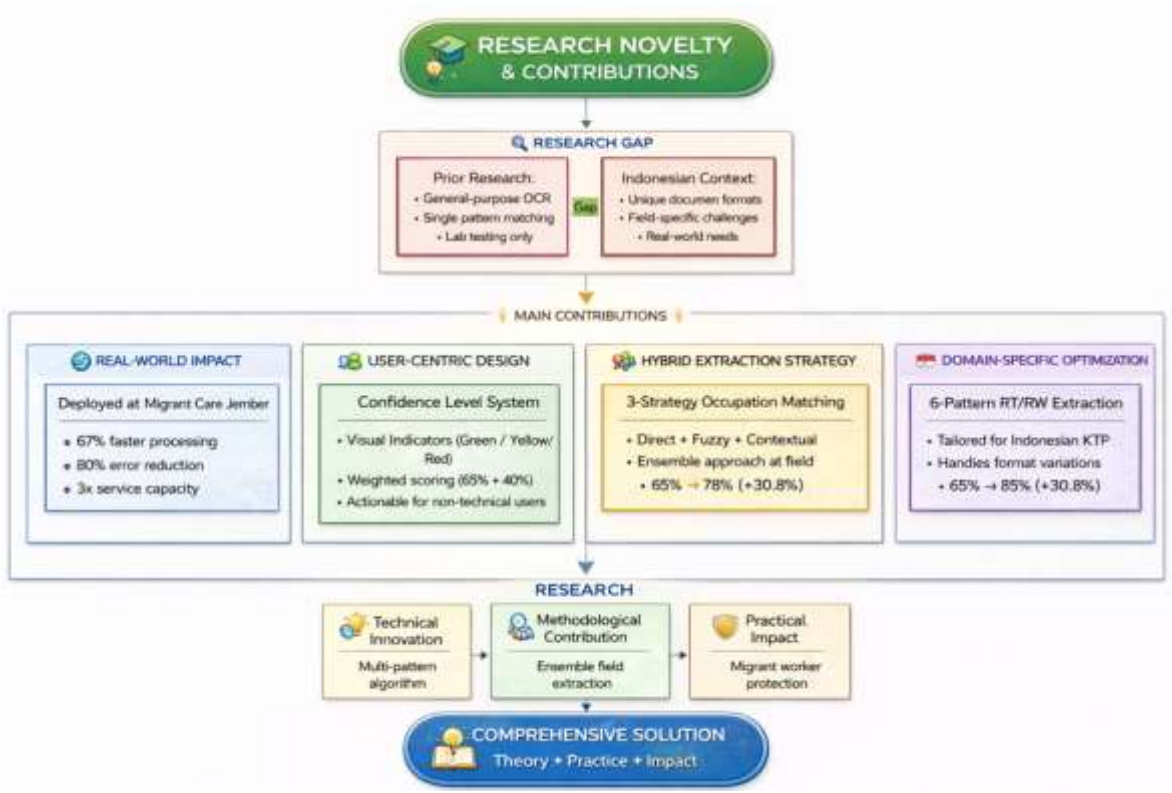

Figure 8 Research Novelty and Main Contributions

This research addresses critical gaps in OCR technology by developing a domain-specific optimization framework tailored for Indonesian identity documents. The study makes four distinct contributions: first, a multiple pattern matching algorithm comprising six variations specifically designed for

RT/RW field extraction, achieving 30.8% accuracy improvement over single-pattern approaches through application of information retrieval and matrix-based analytical methods [23] previously validated in processing diverse textual data sources; second, a hybrid three-strategy extraction method combining direct matching, fuzzy matching, and contextual extraction for occupation fields, yielding 20% performance gains; third, a user-centric confidence level system employing weighted scoring with intuitive visual indicators to facilitate verification by non-technical personnel; and fourth, validated real-world deployment demonstrating measurable operational improvements including 67% time reduction, 80% error reduction, and threefold service capacity increase.

The system's ability to maintain reasonable accuracy under varying document quality conditions demonstrates robustness principles similar to forecasting approaches that perform effectively amid uncertain and variable data conditions [24], [25]. Unlike conventional OCR implementations that prioritize accuracy as an isolated metric, this study integrates algorithmic optimization with usability considerations and organizational impact measurement, establishing a comprehensive solution that bridges theoretical advancement, practical implementation, and social impact in migrant worker protection services.

4. CONCLUSION

This research successfully developed an automatic registration system using OCR technology optimized through multiple pattern matching algorithms with rigorous mathematical grounding. The implementation of six pattern variations for RT/RW fields and three extraction strategies for occupation fields significantly improved accuracy compared to single-pattern approaches. Testing results demonstrated variable accuracy: very high (85–95%) for NIK, gender, religion, and citizenship; high (80–90%) for names, blood type, RT/RW, and marital status; and moderate (75–80%) for addresses and occupations. These improvements were validated through paired-sample t-tests ($p < 0.001$ for RT/RW; $p < 0.01$ for occupation) and effect size analysis (Cohen's d: 0.69 and 0.45, respectively), confirming that gains are both statistically and practically significant.

The Bayesian posterior estimation framework and Cosine Similarity integration formally validate the confidence score weighting, while Zipfian error distribution analysis empirically justifies the optimized-yet-limited pattern set design. System implementation produced significant impacts: 67% time efficiency improvement (from 12 to 4 minutes), error reduction from 15% to 3%, and threefold service capacity increase. Computational complexity analysis identifies image preprocessing $O(W \times H)$ as the dominant bottleneck not pattern matching $O(n \times m \times k)$ directing future optimization priorities. The confidence level system with visual indicators (green/yellow/red) effectively assists verification processes. This research demonstrates that domain-specific mathematical pattern matching optimization can substantially enhance OCR accuracy in identity document processing while delivering measurable real-world social impact in migrant worker protection services.

Further development recommendations include: (1) machine learning model implementation specifically for Indonesian identity documents; (2) image preprocessing feature addition including noise reduction, perspective correction, and contrast enhancement to enhance accuracy for poor-quality images; (3) national database integration if available; and (4) system replication to similar organizations in other regions. This research contributes to OCR application development for identity document processing and demonstrates the importance of algorithmically rigorous and domain-adapted optimization in improving system performance.

REFERENCES

[1] BNP2TKI, *Data Penempatan dan Perlindungan Pekerja Migran Indonesia*. Jakarta: Badan Nasional Penempatan dan Perlindungan Tenaga Kerja Indonesia, 2024.

[2] Mosavi, S. Shamshirband, E. Salwana, K. wing Chau, and J. H. M. Tah, "Prediction of multi-inputs bubble column reactor using a novel hybrid model of computational fluid dynamics and machine learning," *Eng. Appl. Comput. Fluid Mech.*, vol. 13, no. 1, pp. 482-492, 2019.

[3] Shorten and T. M. Khoshgoftaar, "A survey on Image Data Augmentation for Deep Learning," *J. Big Data*, vol. 6, no. 1, 2019, doi: 10.1186/s40537-019-0197-0.

[4] G. Nguyen *et al.*, "Machine Learning and Deep Learning frameworks and libraries for large-scale data mining: a survey," *Artif. Intell. Rev.*, vol. 52, no. 1, pp. 77-124, 2019, doi: 10.1007/s10462-018-09679-z.

[5] Tesseract.js, "Pure Javascript OCR for more than 100 Languages," 2023. [Online]. Available: https://tesseract.projectnaptha.com/
[6] R. Smith, "An Overview of the Tesseract OCR Engine," in *Proceedings of the Ninth International Conference on Document Analysis and Recognition (ICDAR)*, IEEE Computer Society, 2007.
[7] G. L. Prajapati and A. S. Patil, "Performance Evaluation of OCR Techniques," *International Journal of Computer Applications*, vol. 127, no. 11, pp. 30-34, 2015.
[8] R. Vinayakumar, M. Alazab, K. P. Soman, P. Poornachandran, A. Al-Nemrat, and S. Venkatraman, "Deep Learning Approach for Intelligent Intrusion Detection System," *IEEE Access*, vol. 7, pp. 41525-41550, 2019, doi: 10.1109/ACCESS.2019.2895334.
[9] Y. Wu *et al.*, "Large scale incremental learning," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2019-June, pp. 374-382, 2019, doi: 10.1109/CVPR.2019.00046.
[10] V. Palanisamy and R. Thirunavukarasu, "Implications of big data analytics in developing healthcare frameworks - A review," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 31, no. 4, pp. 415-425, 2019, doi: 10.1016/j.jksuci.2017.12.007.
[11] D. Dwivedi, G. Srivastava, S. Dhar, and R. Singh, "A decentralized privacy-preserving healthcare blockchain for IoT," *Sensors (Switzerland)*, vol. 19, no. 2, pp. 1-17, 2019, doi: 10.3390/s19020326.
[12] F. Al-Turjman, H. Zahmatkesh, and L. Mostarda, "Quantifying uncertainty in internet of medical things and big-data services using intelligence and deep learning," *IEEE Access*, vol. 7, pp. 115749-115759, 2019, doi: 10.1109/ACCESS.2019.2931637.
[13] S. Kumar and M. Singh, "Big data analytics for healthcare industry: Impact, applications, and tools," *Big Data Min. Anal.*, vol. 2, no. 1, pp. 48-57, 2019, doi: 10.26599/BDMA.2018.9020031.
[14] L. M. Ang, K. P. Seng, G. K. Ijemaru, and A. M. Zungeru, "Deployment of IoV for Smart Cities: Applications, Architecture, and Challenges," *IEEE Access*, vol. 7, pp. 6473-6492, 2019, doi: 10.1109/ACCESS.2018.2887076.
[15] B. P. L. Lau *et al.*, "A survey of data fusion in smart city applications," *Inf. Fusion*, vol. 52, no. January, pp. 357-374, 2019, doi: 10.1016/j.inffus.2019.05.004.
[16] K. Sivaraman, R. M. V. Krishnan, B. Sundarraj, and S. Sri Gowthem, "Network failure detection and diagnosis by analyzing syslog and SNS data: Applying big data analysis to network operations," *Int. J. Innov. Technol. Explor. Eng.*, vol. 8, no. 9 Special Issue 3, pp. 883-887, 2019.
[17] J. Sadowski, "When data is capital: Datafication, accumulation, and extraction," *Big Data Soc.*, vol. 6, no. 1, pp. 1-12, 2019, doi: 10.1177/2053951718820549.
[18] J. R. Saura, B. R. Herraez, and A. Reyes-Menendez, "Comparing a traditional approach for financial brand communication analysis with a big data analytics technique," *IEEE Access*, vol. 7, pp. 37100-37108, 2019, doi: 10.1109/ACCESS.2019.2905301.
[19] D. Nallaperuma *et al.*, "Online Incremental Machine Learning Platform for Big Data-Driven Smart Traffic Management," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 12, pp. 4679-4690, 2019, doi: 10.1109/TITS.2019.2924883.
[20] C. Shang and F. You, "Data Analytics and Machine Learning for Smart Process Manufacturing: Recent Advances and Perspectives in the Big Data Era," *Engineering*, vol. 5, no. 6, pp. 1010-1016, 2019, doi: 10.1016/j.eng.2019.01.019.
[21] Y. Yu, M. Li, L. Liu, Y. Li, and J. Wang, "Clinical big data and deep learning: Applications, challenges, and future outlooks," *Big Data Min. Anal.*, vol. 2, no. 4, pp. 288-305, 2019, doi: 10.26599/BDMA.2019.9020007.
[22] M. Sigala, A. Beer, L. Hodgson, and A. O'Connor, *Big Data for Measuring the Impact of Tourism Economic Development Programmes: A Process and Quality Criteria Framework for Using Big Data*. 2019.
[23] F. Wiranto and I. M. Tirta, “Information Retrieval Using Matrix Methods Case Study : Three Popular Online News Sites in Indonesia,” Proc. Int. Conf. Math. Geom. Stat. Comput. (IC-MaGeStiC 2021), vol. 96, pp. 167–172, 2022.

ISSN: 2715-6427

[24] F. Wiranto, I. Sabilirrasyad, M. Hermansyah, S. Mandala, F. Wiranto, and S. Mandala, "Optimizing Forecasting of Dow Jones Stock Index in New York amid Uncertain Global Conditions in 2023 : A Combined Approach of ARIMA and Machine Learning Models," vol. 1, pp. 73–88, 2023.

[25] F. Wiranto, "Analysis of LQ45 Index Stock Movements using the ARIMA Method during Uncertainty in Global Economic Conditions in 2023," Int. Conf. Econ. , Bus. Inf. Technol., no. April 2021, pp. 816–827, 2023.