Document Process Automation in an International Logistics Company Using OCR, RPA, and Text Analytics

Nayeli M. Soto Leiva[1], Yair R. Cubas Pecho[2], José A. Taquía Gutiérrez[3], Juan C. Quiroz Flores[4]

[1]20194631@aloe.ulima.edu.pe, [2]20180528@aloe.ulima.edu.pe, [3]jtaquia@ulima.edu.pe, [4]jcquiroz@ulima.edu.pe

[1234]Carrera de Ingeniería Industrial, Universidad de Lima, Perú

Received: July 22, 2025 / Accepted: October 6, 2025 / Published: 5 June, 2026
doi: https://doi.org/10.26439/ciii2025.8661

ABSTRACT—This study aims to optimize the document settlement process in a Peruvian foreign trade logistics company through the implementation of intelligent automation technologies, including robotic process automation (RPA), optical character recognition (OCR), and text analytics. This process, initially characterized by intensive manual tasks, typing errors, and excessive processing times, limited operational performance and generated bottlenecks in logistics operations. Among the main findings, a 49% reduction in average processing time for work orders was observed. During functional validation, the average number of work orders processed per shift was measured. An increase from 4.03 to 6.00 work orders per shift was reported, representing an improvement of 1.97 work orders per shift following the implementation of automation. It is concluded that document automation represents an effective and scalable strategy for improving logistics performance in document processing. As a future goal, automation is planned to be expanded to all administrative areas with repetitive and standardized operational processes.

Index Terms—Data extraction, document processing, optical character recognition (OCR), robotic process automation (RPA), text analytics.

I. Introduction

The logistics sector in Peru has undergone rapid transformation in recent years, driven by digitalization, the growth of e-commerce, and the need to adapt to international standards of operational efficiency. The sector projects an annual growth rate of 7.4%, positioning itself as one of the key economic activities for national competitiveness [1]. However, much of this industry still faces structural challenges, such as the lack of automation in document-related processes, which limits traceability, productivity, and the ability to respond to increased demand.

While leading companies such as DHL and FedEx already operate automated systems that integrate artificial intelligence and document analysis, many Peruvian organizations still rely on manual procedures, achieving on-time delivery (OTD) rates of approximately 85%, below the international benchmark of 95% [1]. This scenario highlights a technological gap that must be addressed through innovative solutions that combine speed, accuracy, and scalability.

In environments where document processes are still largely manual, the lack of automation represents a barrier to achieving effective logistics performance. The absence of intelligent tools to streamline and structure management limits operational responsiveness and hinders business growth in an increasingly demanding market. Faced with this problem, this study aims to optimize the document settlement process through the implementation of intelligent automation technologies.

In response to this problem, this study proposes an intelligent document automation model that integrates robotic process automation (RPA), optical character recognition (OCR), and text analytics. These technologies have proven effective in reducing processing times and error rates in complex document environments [2], [3], [4], [5].

The primary objective of this research is to optimize document flow within logistics processes through an adaptable technological solution validated in a real-world environment. To this end, a system was designed to automatically label files, extract key information, and consolidate data into a structured format. The solution not only aims to improve key performance indicators (KPIs) but also to relieve staff from repetitive tasks, allowing them to focus on higher-value-added activities. From a technological perspective, the proposed solution promotes the adoption of

modern technologies aligned with the sector's digital transformation. From an economic perspective, it contributes to cost reduction by minimizing errors and penalties, thereby improving profitability and operational sustainability.

This was supported by scientific articles, which are classified by typology as shown below:

A. RPA Applied to Document Flows

RPA has proven to be an effective tool for digitizing and structuring repetitive operational workflows. Its use allows reducing manual workload, minimizing human errors, and improving document traceability [2], [5]. In administrative contexts, the combination of RPA with recognition technologies allows automating data extraction from scanned documents and validating them in a structured manner [6]. This type of technology has been widely applied in the banking, government, and logistics sectors, enhancing document processing efficiency while reducing processing times [7].

Furthermore, RPA enables the incorporation of automated validation rules, thereby improving data quality at the source. This approach is particularly useful for repetitive tasks such as data entry, information consolidation, and document verification [6], [8].

B. OCR for Data Extraction

OCR has evolved significantly using machine learning models and deep neural networks. This technology has been optimized to extract information from invoices, contracts, and scanned documents, even when presented in irregular formats or with poor visual quality [1], [4], [9].

Among the main improvements are automatic text slant correction, visual noise removal, and adaptation to diverse fonts and structures [3], [9]. In addition, current models can interpret handwritten, multilingual, and partially damaged content, expanding their applicability in highly demanding environments such as document logistics [10]–[12].

C. RPA and OCR Integration: Operational Synergies

Integrating RPA with OCR enables end-to-end automation of document workflows, from file ingestion to final data consolidation. This approach significantly reduces processing times and human intervention–related errors, particularly in administrative sectors with high document volumes [6], [7], [13].

The combined use of these tools facilitates structured information extraction, automated quality control, and report generation. Recent literature also highlights that their successful implementation depends on adequate management change and the involvement of operational staff in the redesign of automated workflows [14].

This hybrid model is also related to the concept of hyperautomation, understood as the combination of robotic automation with artificial intelligence and machine learning to optimize complex and dynamic processes [15].

D. Text Analytics and Intelligent Document Processing

Intelligent document processing, based on text analytics and machine learning algorithms, enables the automation of tasks beyond simple digitization, by interpreting semantic content, logical structure and contextual information of unstructured documents [12].

This approach enables the automated detection of key entities, semantic relationships among fields, and content-based task prioritization. The use of these tools improves document organization, optimizes traceability, and reduces human intervention in routine review and classification activities [12], [16].

Furthermore, these models have proven to be particularly useful in scenarios involving a high volume of heterogeneous files. Their application has been validated in contexts such as financial auditing, healthcare, and digital logistics, allowing information extraction even from non-standardized structures or in multilingual documents [12], [16], [17].

II. Methodology

A. Bases of the Proposed Model

The proposed solution is based on the integration of RPA, OCR and text analytics. These tools were selected following a systematic review of the scientific literature, which demonstrated their effectiveness in environments with high document volumes. Recent studies confirm that RPA–OCR integration reduces document processing times [7], while semantic processing enhances accuracy in the automatic classification of operational files [8], [12]. In these contexts, the automated extraction of data from structured and semi-structured documents has significantly reduced processing times [18]. Likewise, the use of OCR enhanced with image processing techniques to convert scanned documents into machine-readable data, even under low visual quality conditions, is highlighted [4]. In addition, text analytics–based approaches enable the automatic classification of unstructured files, contributing to improved document organization and operational traceability. [8], [19].

B. Proposed Model

The model proposed in this research addresses the limitations identified in a Peruvian foreign trade logistics company, where a high dependence on manual document processing was observed. This reliance results in delays in customs declarations and directly affects the timely clearance of goods from the seaport. As a solution, a functional intelligent document automation architecture was implemented based on the sequential integration of three technologies: RPA, OCR, and text analytics.

The process begins with the retrieval of files stored in a folder labeled "Work Order" within the company's enterprise resource planning (ERP) system. These documents, primarily in PDF format and lacking a defined structure, are automatically labeled using a Python-based script hosted in a collaborative environment (Google Colab). The script applies text analytics rules to identify document types (e.g., invoices, bills of lading (B/L), policies), thereby addressing the lack of uniform file labeling. [3]. This initial phase improves traceability and facilitates subsequent document validation by the operator in accordance with the corresponding work order.

The labeled documents are subsequently processed by a UiPath-based robot. This robot automatically reads the files using OCR and extracts key fields, such as the taxpayer identification number (RUC in Spanish), dates, amounts, and Incoterms, which are consolidated into a structured template. This step addresses the need to automate data consolidation and mitigate human errors in manual entry [1], [20].

To handle scanned documents or those with low visual quality, the model relies on machine learning–enhanced OCR engines capable of interpreting text under non-ideal conditions [5], [23].

Finally, the extracted data are organized into an Excel spreadsheet using standardized nomenclature, making them ready for validation and subsequent entry into the customs system. This modular and externalized approach enables integration without affecting existing systems, favoring scalability to other document areas.

C. Model Components

1) Diagnosis of the Current Process: The study began with the identification of operational deficiencies in the document settlement process of a logistics company located in Lima, Peru. To this end, qualitative and quantitative analysis tools were applied to characterize the initial situation. In the first stage, semi-structured interviews were conducted with staff from the various involved areas, enabling an understanding of the actual process flow and the identification of its main constraints.

   Subsequently, the current process was mapped, and a Pareto chart was applied to identify the most significant issue within the document clearance area. Based on this finding, an Ishikawa diagram was created to structure the probable causes of the identified problem. These elements were organized into a problem tree to graphically represent the relationship between observed effects and their root causes.

   To validate and quantify the problem, a representative sample of work orders processed under real operating conditions was defined. The initial sample comprised 120 work orders, from which 92 were selected through simple random sampling and used to measure key process indicators.

   Finally, the data obtained were analyzed using discrete-event simulation in Arena software. This analysis enabled an assessment of the process's behavior prior to the intervention, based on three key indicators: average system time, utilization of the human resource responsible for the process (the "settler"), and the number of orders completed per cycle.

2) Design of the Technological Intervention: Based on the findings of the diagnostic phase, a functional architecture integrating three key technologies—RPA, OCR, and text analytics—was designed. This technological proposal was structured into four sequential stages comprising the automated document processing workflow:

   2a) *Preliminary reading and automatic labeling:* A Python-based script was implemented and executed in a collaborative environment (Google Colab), applying a basic OCR engine in conjunction with text analytics rules. This tool enabled the identification of PDF file content associated with each work order and the automatic assignment of labels according to document type (e.g., invoice, policy, manifest).

   2b) *Structured data extraction:* An RPA model developed in UiPath was responsible for directly processing the labeled files. The RPA applied OCR with advanced parameters and identified key fields using predefined semantic rules—such as invoice number, supplier tax identification number, issue date, and Incoterms—as illustrated in Fig. 1.

   2c) *Data validation and organization:* Once the extraction was complete, the RPA automatically organized the data into an Excel spreadsheet, enabling rapid access to the document information. This structuring aims to significantly reduce the time required for manual input into customs document declaration platforms, while also minimizing human error in data entry.

3) Model Validation: The effectiveness of the proposed model was validated through a comparative simulation between the baseline scenario (without improvement) and the intervened scenario. The simulation results were analyzed using the Output Analyzer platform. The results aim to demonstrate the value of the proposed approach by analyzing the number of work orders processed per shift, the average time a work order remains in the system, and the utilization of operational resources, specifically the document settlement operator.

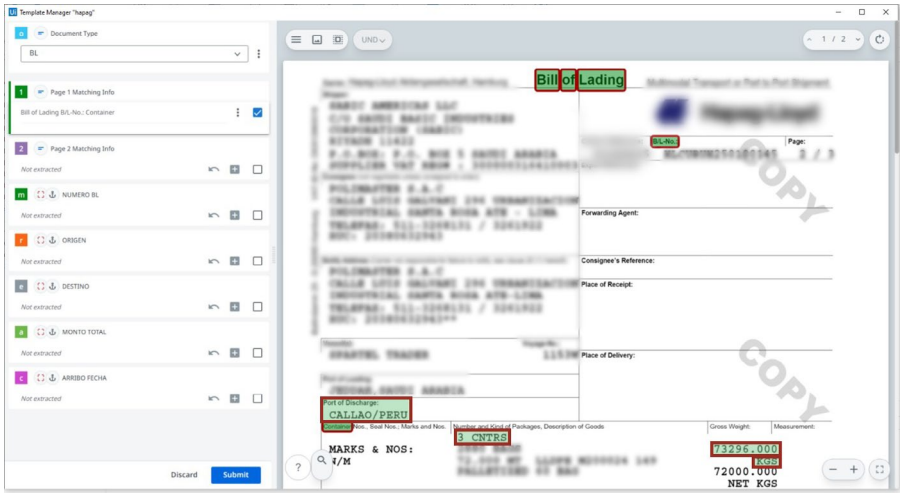

Fig. 1. Data extraction from a bill of lading (B/L) document.

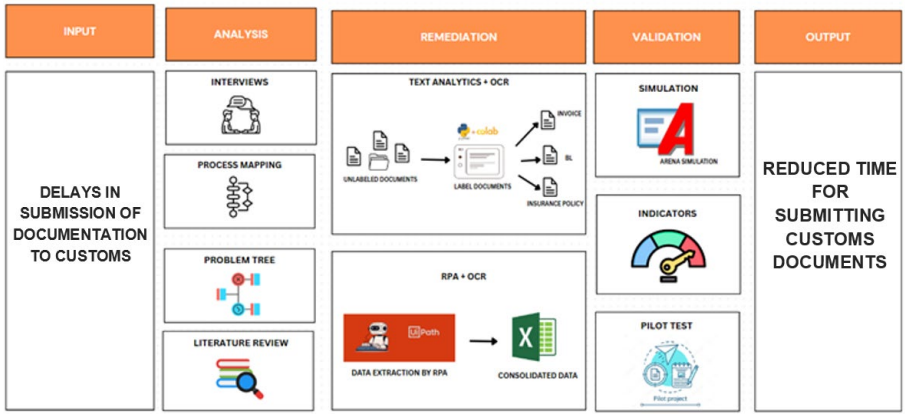

Fig. 2. Model of RPA and OCR implementation.

Fig. 2 illustrates the proposed conceptual model, highlighting the expected outcomes of implementing the defined processes and tools.

III. Results

The proposed model was validated through simulations in Arena software, as shown in Fig. 5, and a pilot test in a company in the foreign trade logistics sector. For the simulation in Arena, 479 replications were run to ensure statistical reliability with a 94% confidence level. The results showed substantial improvements across three key indicators: the number of work orders processed per 8-h shift, the average system time per work order, and the utilization rate of the primary operational resource (liquidator), as summarized in Table I.

In the baseline scenario, the system processed an average of 5.45 records per 8-h shift, with a 94% confidence interval of [5.39, 5.52]. After implementing the RPA–OCR–text analytics solution, this indicator increased to 7.39 records per shift, with a 94% confidence interval of [7.35, 7.42], representing a 35% productivity improvement without additional staffing. The average time a document remained in the system decreased from 123 min in the baseline scenario to 62.7 min after automation, representing a 49% reduction. This improvement directly impacts operational agility and the timely availability of information for subsequent processes. Regarding resource utilization, the liquidator's utilization slightly decreased from 99.7% to 89.2%. Although this reduction is moderate, it is significant, as a higher number of documents were processed within the same shift while maintaining—and slightly reducing—the staff workload. These results indicate that the automated system enables a more fluid and efficient workflow by optimizing the liquidator's time and improving their productivity, as shown in Fig. 3.

Statistical comparison using the paired t-test confirmed that the changes achieved with the proposed solution were not random. For all performance indicators, the confidence intervals for the mean differences did not include zero, confirming that the improvements are statistically significant and attributable to the implemented redesign, as shown in Fig. 4.

Additionally, a pilot test was initiated in May 2025, during which partial operation of the document automation tool was implemented. In this context, the monthly evolution of the average processing time per order is presented in Fig. 6. During the first eight months analyzed, processing times remained high, ranging from 105 to 130 min, reflecting the operational burden of the manual process and remaining within the parameters estimated by the Output Analyzer. However, starting in May—coinciding with the technological intervention—a progressive improvement became evident. In June and July, despite minor fluctuations, processing times remained below 71 min, consolidating a sustained reduction compared to the previous period.

This behavior suggests that the automated system enabled a more streamlined workflow, allowing staff to validate data more efficiently without the need to perform repetitive manual tasks. Although the values still reflect a learning curve for operators, the trend shows that the technological implementation effectively contributed to reducing bottlenecks, improving productivity, and reducing human errors in the customs documentation process.

IV. Discussion

The behavior observed in the validation phase suggests that the automated system enabled a smoother workflow, allowing staff to validate data more efficiently while reducing manual, repetitive tasks. Although the results still reflect a learning curve for operators, the trend indicates that the technological implementation effectively contributed to reducing bottlenecks, improving productivity, while reducing human errors in the customs documentation process.

The results obtained from the simulation and pilot test confirm that intelligent automation applied to the document clearance process is a viable and effective strategy for addressing the main operational issues arising from manual information processing.

The integration of RPA, OCR, and text analytics enabled the redesign of the document flow by eliminating repetitive tasks, reducing human errors, and significantly shortening execution times. These findings are consistent with prior studies that highlight the benefits of robotic automation in improving efficiency, traceability, and data accuracy in administrative processes [2], [5], [6], [7].

In addition, the use of advanced OCR optimized through machine learning was critical for interpreting scanned documents with poor visual quality. This capability has been documented in recent studies that emphasize the robustness of current OCR engines when handling non-editable, degraded, or multi-font documents [4], [9], [11].

The inclusion of text analytics techniques addressed one of the primary limitations identified during the initial diagnosis: the lack of standardized file nomenclature. These techniques enabled automatic document classification, facilitated file retrieval, and improved overall traceability [12], [16], [17].

From a digital transformation perspective, this model represents a step toward hyperautomation, understood as the integration of RPA with artificial intelligence and semantic processing to optimize complex processes [15].

Likewise, the quantitative results showed a 35% increase in the number of orders processed per shift and

TABLE I

Comparison of Indicators Before and After Improvement

Indicators	As is	To be
Number of orders processed	5.45	7.39
Average time (min) of the work order in the system	123	62.7
% utilization of the liquidator	99.7%	89.2%

Prepared by authors.

Fig. 3. Work order comparison.

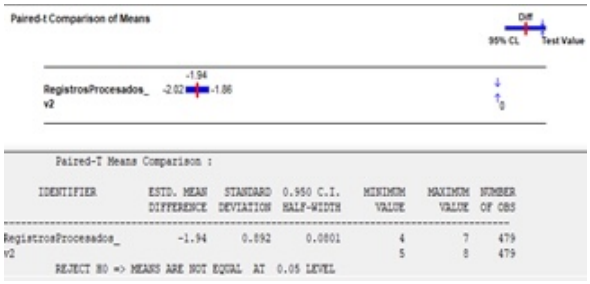

Fig. 4. Paired t-test validation. Authors' elaboration.

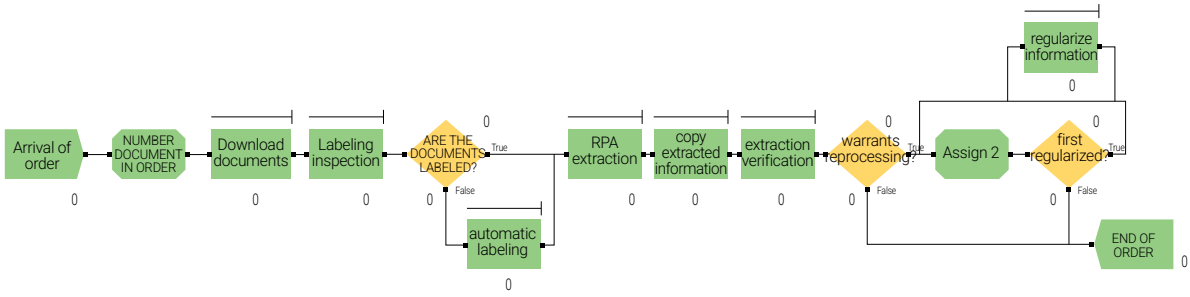

Fig. 5. Flowchart in Arena of the improvement proposal. Authors' elaboration.

a 49% reduction in the average processing time per order. These indicators demonstrate not only improved operational efficiency but also more effective utilization of human resources, as staff workload was reduced without compromising productivity—and, in fact, with a measurable improvement.

Although an operational learning curve was identified, the progressive trend observed after the intervention—as evidenced by the pilot test—validates the effectiveness of the proposed technological architecture and supports its scalability to other areas of the organization.

Furthermore, this model can be adapted to other administrative sectors with similar characteristics —such as standardized and repetitive workflows, high document volumes, and traceability requirements—thereby reinforcing its applicability as a comprehensive solution within institutional modernization processes.

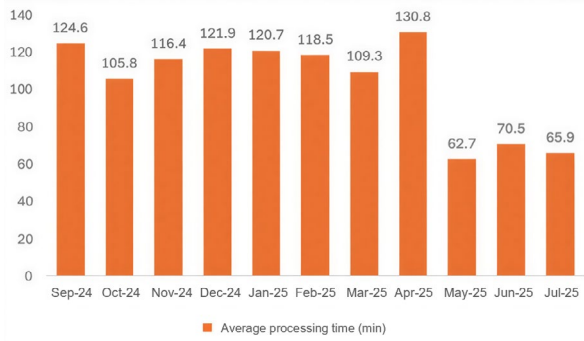

Fig. 6. Average processing times for a work order.

V. Conclusions

This research demonstrates that intelligent automation, applied to the document settlement process in logistics operations, is an effective approach for mitigating the primary operational issues associated with manual information processing. Using a model comprised of RPA, OCR, and text analytics, the work order processing flow was redesigned, eliminating repetitive tasks, reducing human errors, while significantly shortening execution times.

The model was validated through both simulation and real-world pilot testing, achieving a 35% increase in the number of orders processed per cycle and a 49% reduction in average order processing time. These results demonstrate that the integration of automation technologies not only enhances human resource productivity but also improves the traceability and accuracy of document records, which is critical in customs-related operations.

Furthermore, the use of text analytics tools helped overcome initial limitations such as the lack of standardized file nomenclature, thereby facilitating document access and classification prior to data extraction. RPA, in turn, proved effective in structuring, validating, and consolidating critical information, even from non-editable or low-quality documents, through its integration with advanced OCR.

Finally, the results observed in the months following the technological intervention indicate a sustained improvement trend, despite staff remaining in a learning-curve phase. This reinforces the viability of the proposed model and supports its scalability to other internal processes. Consequently, the progressive adoption of intelligent automation–based solutions represents a fundamental step toward modernizing logistics services in a highly demanding and competitive environment.

References

[1] Ministerio de Economía y Finanzas (MEF), "Marco Macroeconómico Multianual 2024–2027" ["Multiannual Macroeconomic Framework 2024–2027"], Gob.pe, 2024. [Online]. Available: https://www.gob.pe/institucion/mef/informes-publicaciones/5603986-marco-macroeconomico-multianual-2024-2027

[2] C. A. Bermúdez Irreño, "RPA – automatización robótica de procesos: Una revisión de la literatura"

["RPA – robotic process automation: A review of the literature"], *Rev. Ing. Mat. Cienc. Inf.*, vol. 8, no. 15, pp. 111–122, 2021, doi: https://doi.org/10.21017/rimci.2021.v8.n15.a97

[3] T. Saout, F. Lardeux, and F. Saubion, "An overview of data extraction from invoices," *IEEE Access*, vol. 12, pp. 19872–19886, Jan. 2024, doi: https://doi.org/10.1109/ACCESS.2024.3360528

[4] A. Haleem, M. Javaid, R. P. Singh, S. Rab, and R. Suman, "Hyperautomation for the enhancement of automation in industries," *Sens. Int.*, vol. 2, 100124, Aug. 2021, doi: https://doi.org/10.1016/j.sintl.2021.100124

[5] Instituto Nacional de Estadística e Informática, *"PBI de las actividades económicas por años* [GDP of economic activities by year]," Instituto Nacional de Estadística e Informática, 2023. [Online]. Available: https://www.inei.gob.pe/estadisticas/indice-tematico/pbi-de-las-actividades-economicas-por-anos-9096

[6] A. Waqar *et al.*, "Assessment of barriers to robotic process automation (RPA) implementation in safety management of tall buildings," *Buildings*, vol. 13, no. 7, p. 1663, Jun. 2023, doi: https://doi.org/10.3390/buildings13071663

[7] K. A. Saavedra Mera, B. M. Quiñonez Cabeza, A. H. Quiñonez Klinger, and V. J. Sarango Romero, "La digitalización de la cadena de suministro: Un impulso innovador para la eficiencia logística en Ecuador [Supply chain digitalization: An innovative boost for logistics efficiency in Ecuador]," *Código Cient. Rev. Investig.*, vol. 4, no. 2, pp. 210–224, Dec. 2023, doi: https://doi.org/10.55813/gaea/ccri/v4/n2/238

[8] A. A. Manjunath *et al.*, "Automated invoice data extraction using image processing," *IAES Int. J. Artif. Intell.*, vol. 12, no. 2, pp. 514–521, Jun. 2023, doi: https://doi.org/10.11591/ijai.v12.i2.pp514-521

[9] S. A. Francis and M. Sangeetha, "A comparison study on optical character recognition models in mathematical equations and in any language," *Results Control Optim.*, vol. 18, no. 1, Art. no. 100532, Mar. 2025, doi: https://doi.org/10.1016/j.rico.2025.100532

[10] A. Deekshith, "Advances in natural language processing: A survey of techniques," *Int. J. Innov. Eng. Res. Technol.*, vol. 8, no. 3, pp. 74–83, Oct. 2024, doi: https://doi.org/10.26662/ijiert.v8i3.pp74-83

[11] J. Villena Toro, A. Wiberg, and M. Tarkian, "Optical character recognition on engineering drawings to achieve automation in production quality control," *Front. Manuf. Technol.*, vol. 3, 1154132, Mar. 2023, doi: https://doi.org/10.3389/fmtec.2023.1154132

[12] J. Kokina, S. Blanchette, T. H. Davenport, and D. Pachamanova, "Challenges and opportunities for artificial intelligence in auditing: Evidence from the field," *Int. J. Account. Inf. Syst.*, vol. 56, 100734, Dec. 2025, doi: https://doi.org/10.1016/j.accinf.2025.100734

[13] D. C. Villarreal Meza, M. G. Cevallos Vizuete, D. C. Arias Portalanza, and K. A. Moya Palacios, "Optimización de los procesos de logística, su mejora y satisfacción al cliente [Optimization of logistics processes, its improvement and customer satisfaction]," *Conciencia Digital*, vol. 5, no. 1.3, pp. 216–233, Mar. 2022, doi: https://doi.org/10.33262/concienciadigital.v5i1.3.2137

[14] L. Isaza and K. Cepa, "Automation and augmentation: A process study of how robotization shapes tasks of operational employees," *Eur. Manag. J.*, Dec. 2024, doi: https://doi.org/10.1016/j.emj.2024.11.010

[15] J. Ribeiro, R. Lima, T. Eckhardt, and S. Paiva, "Robotic process automation and artificial intelligence in industry 4.0: A literature review," *Procedia Comput. Sci.*, vol. 181, pp. 51–58, Jan. 2021, doi: https://doi.org/10.1016/j.procs.2021.01.104

[16] K. Soeny, G. Pandey, U. Gupta, A. Trivedi, M. Gupta, and G. Agarwal, "Attended robotic process automation of prescriptions' digitization," *Smart Health*, vol. 20, 100189, Apr. 2021, doi: https://doi.org/10.1016/j.smhl.2021.100189

[17] C. Flechsig, F. Anslinger, and R. Lasch, "Robotic process automation in purchasing and supply management: A multiple case study on potentials, barriers, and implementation," *J. Purch. Supply Manag.*, vol. 28, no. 1, 100718, Jan. 2022, doi: https://doi.org/10.1016/j.pursup.2021.100718

[18] S. İ. Omurca, E. Ekinci, S. Sevim, E. B. Edinç, S. Eken, and A. Sayar, "A document image classification system fusing deep and machine learning models," *Appl. Intell.*, vol. 53, no. 12, pp. 15295–15310, Nov. 2022, doi: https://doi.org/10.1007/s10489-022-04306-5

[19] N. Gal-Nadasan, V. Stoicu-Tivadar, E. Gal-Nadasan, and A. R. Dinu, "Robotic process automation based data extraction from handwritten medical forms," *Stud. Health Technol. Inform.*, vol. 309, pp. 68–72, Oct. 2023, doi: https://doi.org/10.3233/SHTI230741

[20] S.-H. Kim, "Development of evaluation criteria for robotic process automation (RPA) solution selection," *Electronics*, vol. 12, no. 4, p. 986, Feb. 2023, doi: https://doi.org/10.3390/electronics12040986

[21] M. Borkowski, W. Fdhila, M. Nardelli, S. Rinderle-Ma, and S. Schulte, "Event-based failure prediction in distributed business processes," *Inf. Syst.*, vol. 81, pp. 220–235, Mar. 2019, doi: https://doi.org/10.1016/j.is.2017.12.005

[22] A. S. Villar and N. Khan, "Robotic process automation in banking industry: A case study on Deutsche Bank," *J. Bank. Financ. Technol.*, vol. 5, no. 1, pp. 71–86, May 2021, doi: https://doi.org/10.1007/s42786-021-00030-9

[23] Mordor Intelligence, "Peru freight and logistics market size & share analysis: Growth trends and forecast (2025–2030)," Mordor Intelligence, 2025. [Online]. Available: https://www.mordorintelligence.com/industry-reports/peru-freight-and-logistics-market