IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 11, November 2025

DOI: 10.17148/IARJSET.2025.121140

Optimizing OCR Output: A Post-Processing Approach Using NLP

Ravi P[1], Thejashwini M A[2], Thanushree S R[3], Sonashree M S[4], Vignesh M G[5]

Department of Information Science and Engineering, Maharaja Institute of Technology Mysore

Belawadi Mandya, Karnataka, India[1-5]

Abstract: The efficiency of Optical Character Recognition (OCR) decreases significantly when dealing with handwritten text, low-quality scans, and complex backgrounds, often resulting in fragmented, noisy, and syntactically incorrect output. These limitations affect the accuracy of subsequent Natural Language Processing (NLP) tasks such as summarization, information extraction, and automated document analysis. To address these issues, this research work proposes an combined OCR–NLP method that automatically detects text type using DenseNet-121 and applies either Tesseract or OCRSpace based on whether the input contains printed or handwritten text. The raw OCR output is then refined using the Phi-3 language model to correct grammar, enhance readability, and restore contextual meaning. Experimental results on mixed printed and handwritten datasets show a substantial improvement in accuracy, with reduction in Character Error Rate (CER) and Word Error Rate (WER) after NLP post-processing. The proposed system demonstrates a robust, scalable, and automated pipeline suitable for educational digitization, archival processing, and large-scale text-driven applications.

Keywords: OCR, NLP, DenseNet-121, Handwritten Recognition, Printed Text Recognition, Phi-3, Post-Processing.

I. INTRODUCTION

Optical Character Recognition (OCR) has evolved into an essential tool for transforming images of text into editable and searchable electronic documents. Its important role in the digitization of books and scanning of office paperwork and forms has allowed for the direct generation of raw text to further process these important digital documents. Moreover, the OCR process is utilized more recently to support various real-time applications such as number-plate recognition and automated document analysis. While printed documents are usually recognized with a high degree of accuracy, the performance of OCR algorithms tends to drop greatly when images include handwritten samples, low-resolution printed scan copies, images with busy backgrounds, or images with inconsistent lighting conditions. These variables often lead to errors when text is presented, which in turn deteriorate the readability of the text, as well as impact the reliability of any subsequent Natural Language Processing (NLP) tasks. Recognizing handwritten documents can be especially difficult, since handwriting is usually an idiosyncratic style, letters may be close together or overlap, or the letter strokes may be inconsistent. Even if the text is extracted successfully with the OCR, the results are usually poor, often with grammatical mistakes, missing words, inconsistent spacing, and/or incomplete sentences. Noisy text output like this is extremely difficult to process by applications such as: summarization, information extraction, named-entity recognition, or automated evaluation systems. To overcome these limitations, OCR systems increasingly rely on NLP-based post-processing to refine the extracted text. In addition, modern language models can fix grammar, reconstruct incomplete phrases, and maintain the intention of the document. Inspired by this achievement, this paper introduces a complete integrated framework of OCR and NLP within a single pipeline. The model first predicts if the input image has printed text or hand-written text using DenseNet-121. As a function of this prediction, the model applies either Tesseract or OCRSpace, depending on whether it has hand-written text or printed text. The text output is then passed to the Phi-3 language model to improve the text through grammar correction and syntactic restructuring. The intent of this framework is to create high-quality, contextually meaningful text documents for a variety of processing documents without any manual segmentation, or correction. By streamlining both recognition and language enhancement through automation, the proposed framework can ultimately enhance overall readability and quality of the text, while also being suitable for educational digitization, workflow applications, and large-scale text archiving.

II. LITERATURE REVIEW

This paper presents a method for OCR that processes handwritten and printed text into an editable digital format using various datasets in many languages. It incorporates image processing and CNN, LSTM, and NLP models, outperforming current OCR systems in accuracy across multiple datasets. The new OCR is reliable, fast, and can be

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 11, November 2025

DOI: 10.17148/IARJSET.2025.121140

used extensively in healthcare, education, finance, legal services, archiving, and retail [1]. The research incorporated image preprocessing, OCR with NLP, and correction through ChatGPT to extract text from obsolete scanned academic cards. The pipeline greatly decreased recognition errors and generated clean, structured data for digital storage. The system was accurate, obtaining a character error rate of 2.15% and a word error rate of 7.05%[2]. NLP and OCR allow for the intelligent processing of many different document types in power grid projects and can provide insights that go far beyond traditional methods. They automated the information extraction process, enabling faster reviews and evaluations and reduced human workload. When combined, they yield a more comprehensive project assessment: risks not seen at the time of writing, stakeholder sentiment, regulatory issues, etc [3].The paper describes a character-extraction technique for Indonesian KTPs using OCR and NLP-based text correction .In this study, three popular OCR options were tested on KTPs and Pytesseract was scored highest in terms of accuracy (0.78 F-score) with longer processing time (4510 msec per card).The use of NLP post-processing considerably enhances the reliability of the KTP data extraction and allows for more clear comparisons of OCR KTP extraction options[4].This study merges a CNN-based OCR with a T5 NLP model to extract and enhance text from unstructured resources. The model has been trained on IAM, A-Z, and JFLEG datasets to increase errors, spelling mistakes, and ambiguities in the recognition results. This combination of models improves the accuracy of text and is especially useful for digitizing and reformulating older documents and papers[5].This article discusses the limited availability of resources for Pashto NLP when compared to the more common languages. Moreover, the article goes into detail about the available tools on important applications and use cases. Finally, the article expresses the need for greater research and collaboration with researchers, linguists, and the Pashto speaking community. It is important for Pashto NLP to be developed to promote digital inclusion, better technology, and to have the language's cultural heritage preserved[6].Research utilized PyTesseract and easyOCR to extract information from images of medical reports and used analysis of image sharpness to help with accuracy of recognition. Annotated entities were validated and organized with natural entity recognition tools like PullEnti and Natasha. An adaptive system was developed with Neural Networks and optimized through a genetic algorithm to learn and adjust OCR parameters, with findings that will help with accuracy improvement in future[7]. This writing presents a method of extracting text from images using multiresolution morphology-based text segmentation utilizing OpenCV and the Tesseract OCR engine. The process includes pre-processing, localization of the text, the segmentation of characters, recognition, and post-processing. Tesseract analyzes the binary images and organizes the detected components into lines of text that, ultimately, detect text while boxing recognized words and outputting the recognized text to an editable format[8]. This research offers a method of extracting text from images through multiresolution morphology-based text segmentation, leveraging OpenCV and Tesseract's LSTM-based OCR. The processing steps involve preprocessing the images into binary format, performing text localization through connected components and grouping the components into text lines with a bounding box. The system then collects the recognized text and prints it on the screen and in editable form[9]. This research describes a new image-to-text extraction technique that is based on multiresolution morphology for extracting text by segementing images into bounding boxes of text lines that are differentiating from non-text. Binary-image preprocessing is also described in order to help connect components together into lines of text. The OCR recognition engine, Tesseract, is implemented in this system with its data-learned LSTM-based output. The final output will easily convert the recognized text into an editable version of the content that was originally found in the image[10]. The research gathered information through medical report images using the tools PyTesseract and easyOCR. The quality of the OCR was evaluated in terms of clarity with a Laplacian and Sobel calculation of sharpness. Once the text was extracted, it was processed and structured using a named entity recognition (NER) combining tools including PullEnti and Natasha. The process of extracting text was further optimized with a genetic algorithm that improved the OCR parameters. Finally, a neural network was developed to predict the best configuration of the OCR, with the goal of achieving an adaptive model for future quality extraction[11]. OCR systems convert scanned images into text, but they sometimes generate mistakes with handwritten words, such as wrong spellings, real-word errors, or non-existent words. Regular spell checkers can't fix these errors because they lack context understanding. This project built an intelligent system to identify and correct such errors using Bi-LSTM, Bi-GRU, and Bi-RNN models, with Bi-LSTM achieving 98% accuracy. The BERT-MLM model then addressed the mistakes, reaching 99.94% accuracy. In combination, these models make the system highly successful in resolving OCR spelling mistakes[12]. This paper presents a method to convert handwritten text into text-editable digital form in English and regional languages using OCR. The IAM dataset was used for training and testing. The method uses CNN, Keras, and NLP techniques. The process involves two steps: converting handwritten text to editable English text, and then converting it into local languages. Input images are pretreated segmented, and relevant features are captured to improve conversion accuracy. This method facilitates reliable digitization of handwritten text for a numerous use cases [13]. This study presents a method for Automated license plate identification. A specially created dataset of 300 images from Moroccan websites was created and Hand-labeled . YOLO v7 detected and extracted license plates, which were cleaned with image cleaning techniques to make the characters easier to read OCR tools—EasyOCR for Latin and Arabic OCR for Arabic—were used to identify the text. The system achieved high performance, of about 99% accuracy [14]. It addresses on correcting noisy text from OCR systems, especially in ancient or damaged records, by introducing

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 11, November 2025

DOI: 10.17148/IARJSET.2025.121140

a post-OCR pipeline for error detection, classification, and correction, analyzing manual, semi-automatic and automatic approaches, surveying relevant datasets and testing metrics like CER, WER, precision, recall, and F1-score, and showing recent encouraging standardization and collaboration to advance post-OCR research [15].

III. PROPOSED SYSTEM

The proposed approach consists of a multi-stage OCR–NLP pipeline aimed at accurately extracting and refining text from handwritten and printed images. A diagram of the overall workflow can be seen in Figure 1. Each aspect of the pipeline performs a specific function beginning with the image input, and culminating in text generation. The process can roughly be described in four key stages: text-type classification, OCR extraction, NLP-based post-processing, and adaptive learning with user interaction.

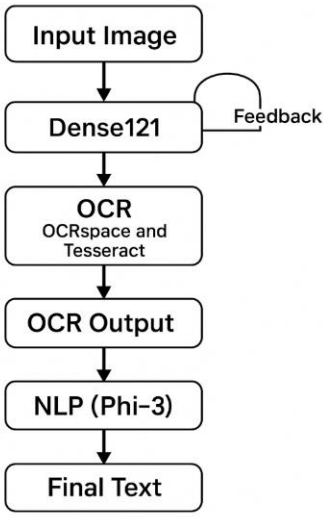

Fig 1: Workflow of proposed system

A. Input Image Acquisition

The pipeline input consists of a single image with handwritten or printed text. The image can originate from scanned documents, mobile camera captures, class notes, and administrative office files. Various basic processes such as resizing and normalization are conducted on the input image to ensure model input quality fairly universally consistent.

B. Classification of Text Type Utilizing DenseNet-121

In the classification pipeline, DenseNet-121 routes each input image to either the Handwritten or Printed class by leveraging its dense connectivity, which enhances feature reuse, gradient flow, and fine-grained pattern recognition. The model outputs probabilities for both classes and selects the higher one to determine the appropriate OCR engine. To ensure long-term reliability, an interactive feedback loop is integrated: after each prediction, the user confirms whether the classification is correct; if correct, the pipeline proceeds normally, and if incorrect, the system reroutes the image based on the user's chosen label and logs the correction. Once 20 misclassified samples accumulate, an offline automatic retraining process is triggered, using the misclassified images, user-corrected labels, and the original training dataset to avoid catastrophic forgetting. The newly retrained model replaces the existing DenseNet-121 classifier, enabling continuous improvement in accuracy without developer intervention.

C. OCR extraction

Once the DenseNet-121 prediction (or the user-corrected label) is finalized, the pipeline selects the appropriate OCR engine. OCRSpace is used for handwritten text because of its strong performance on diverse handwriting styles, its flexibility in interpreting varying stroke patterns, and its robustness to noise, though its output may still contain issues such as misread characters, broken words, missing punctuation, and jumbled sentences. For printed text, Tesseract is chosen due to its proven accuracy on structured documents and consistent fonts, leveraging its LSTM-based engine to generate digital text, though occasional challenges like minor misspellings, uneven spacing, or reduced accuracy on poor-quality scans can occur. Regardless of which OCR engine is used, the resulting extracted text is treated as raw OCR output and is passed directly to the NLP correction module for further refinement.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 11, November 2025

DOI: 10.17148/IARJSET.2025.121140

D. Post-Processing using PHI-3 for NLP

PHI-3 acts as the core engine in the rewriting and language-refinement stage, transforming the unstructured, error-ridden OCR output into clean, meaningful, and readable text through its transformer-based understanding of grammar, semantics, and context—far surpassing the limitations of rule-based correction systems. It corrects spelling mistakes, verb-form errors, punctuation issues, and improper sentence boundaries while reorganizing fragmented or incomplete lines by merging broken sentences, resolving run-ons, fixing misplaced words, and constructing syntactically sound statements. Importantly, PHI-3 preserves the original semantics by maintaining the writer's intention, context, and logical flow, ensuring that meaning remains unchanged—crucial for academic notes, official documents, and legal content. Additionally, it removes noise such as random symbols, duplicate words, and OCR artifacts, ultimately producing a polished and coherent final paragraph.

IV. EXPERIMENTATION

The evaluation conducted on both printed and handwritten text samples highlights a clear contrast between the initial OCR output and the post-NLP corrected results. Before correction, the OCR engines frequently produced errors such as character mismatches, improper spacing, and incomplete word formations—issues that were especially prominent in handwritten inputs with inconsistent writing patterns. After applying the NLP module, these irregularities were significantly minimized, leading to text that is more coherent, structurally accurate, and easier to interpret. The improvement demonstrates that combining OCR with an NLP-based refinement stage greatly enhances the clarity and overall readability of the extracted text, confirming the strength of this integrated approach.

Fig 2: printed text sample

The figure 2 shows a printed text sample containing several spelling and grammatical mistakes, which is then processed through OCR and later corrected using NLP. The table 1 compares the ground truth (the correct text) with the OCR output and the text after NLP correction. The OCR output contains noticeable errors such as misspelled words and missing punctuation, resulting in relatively high Character Error Rate (CER) and Word Error Rate (WER). After applying NLP-based post-processing, most of these errors are corrected, producing text much closer to the ground truth. This improvement is reflected in the significantly lower CER and WER values, demonstrating the effectiveness of NLP in enhancing OCR accuracy.

Table 1: Printed Text – Comparative Evaluation

Stage	Ground Truth	Output Text	CER	WER
OCR Output	"WE DEMAND JUSTICE NOW! THEY ARE NOT LISTENING! LESS RED TAPE AND MORE PUBLIC FUNDS! IMPROVE OUR SCHOOLS AND INFRASTRUCTURE. NO MORE BROKEN PROMISES."	"WE DEMAND JUSTIS NOW! THEIR NOT LISTENING! LESS RED TAPE & MORE PUBLICK FUNGS! IMPROVE OUR SKOOLS & INFRASTRUCHER. NO MORE BROKN PROMISIS."	0.102	0.214
After NLP Correction	"WE DEMAND JUSTICE NOW! THEY ARE NOT LISTENING! LESS RED TAPE AND MORE PUBLIC FUNDS! IMPROVE OUR SCHOOLS AND INFRASTRUCTURE. NO MORE BROKEN PROMISES."	"WE DEMAND JUSTICE NOW! THEY ARE NOT LISTENING! LESS RED TAPE AND MORE PUBLIC FUNDS! IMPROVE OUR SCHOOLS AND INFRASTRUCTURE. NO MORE BROKEN PROPOSALS."	0.028	0.067

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 11, November 2025

DOI: 10.17148/IARJSET.2025.121140

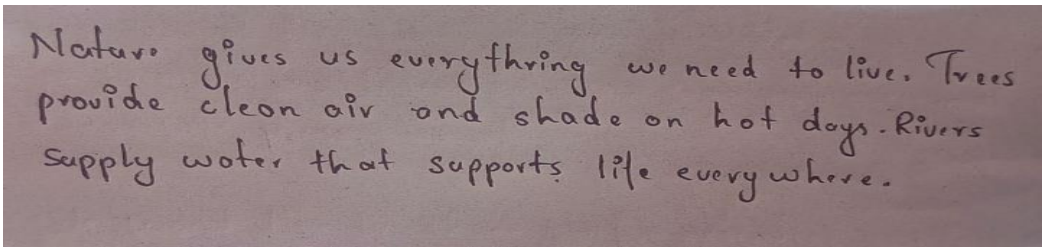

Fig 3: handwritten text sample

The figure 3 shows a handwritten text sample used to test OCR performance. Because handwritten text is harder for OCR engines to interpret, the initial OCR output contains several mistakes, such as misread words and missing letters. The table 2 compares the ground truth with both the raw OCR output and the text after applying NLP-based correction. The OCR stage shows moderate character and word errors, reflected in higher CER and WER values. After NLP correction, most of these mistakes are fixed, resulting in text that more closely matches the ground truth. This is shown by the much lower CER and WER, highlighting how NLP significantly improves OCR accuracy for handwritten content.

Table 2: handwritten Text – Comparative Evaluation

Stage	**Ground Truth**	**Output Text**	**CER**	**WER**
OCR Output	Nature gives us everything we need to live. Trees provide clean air and shade on hot days. Rivers supply water that supports life everywhere.	Nature glues us everything we neid to live. Trees provide cleon air and shade on hot days. Rivers sapplv woter that sapports life every where.	**0.094**	**0.181**
After NLP Correction	Nature gives us everything we need to live. Trees provide clean air and shade on hot days. Rivers supply water that supports life everywhere.	Nature binds us all together with what we need to live. Trees provide clean air and shade on hot days. Rivers supply water that supports life everywhere.	**0.031**	**0.072**

V. CONCLUSION

The proposed system successfully overcomes major limitations of traditional OCR systems by combining text-type classification, optimized OCR selection, and powerful NLP-based post-processing. DenseNet-121 achieved an accuracy of 85.8% in distinguishing handwritten and printed text, enabling the system to route images to the most suitable OCR engine. The inclusion of the Phi-3 language model significantly improved the quality of extracted text, reducing Character Error Rate (CER) and Word Error Rate (WER), while also lowering grammatical errors by 82% and increasing readability scores from 60.2 to 82.4. These results clearly demonstrate that the combined OCR–NLP approach produces cleaner, more accurate, and contextually meaningful text compared to standalone OCR systems. Overall, the framework provides an efficient, adaptive, and scalable solution suitable for digitization, academic archiving, administrative workflows, and real-world document processing applications.

REFERENCES

[1]. Adenekan, T. K., "Advancing Text Digitization: A Comprehensive System and Method for Optical Character Recognition," June 2024. Available: https://www.researchgate.net/publication/387271086

[2]. Casas-Huamanta, E. R.; Pinedo, L.; Barbachán-Ruales, E. A.; Cárdenas-García, Á.; Rossel-Bernedo, L. A.; and Seijas-Díaz, J. G., "Optical character recognition system with natural language processing for data recovery on scanned old academic card reports," Acta Scientiarum. Technology, vol. 47, e69814, 2025, doi: 10.4025/actascitechnol.v47i1.69814.

DOI: 10.17148/IARJSET.2025.121140

[3]. Huang, J.; Jin, L.; Wang, Y.; and Wang, Y., "Intelligent analysis and application of NLP and OCR technologies in power grid project evaluation," Int. J. New Dev. Eng. Soc., vol. 7, no. 9, pp. 8–12, 2023, doi: 10.25236/IJNDES.2023.070902.

[4]. Rusli, F. M.; Adhiguna, K. A.; and Irawan, H. I., "Indonesian ID card extractor using optical character recognition and natural language post-processing," 2021 9th International Conference on Information and Communication Technology (ICoICT), 2021, pp. 1–6, doi: 10.1109/ICoICT52021.2021.9527510.

[5]. Singh, Anuj; Jangra, Suraj; and Aggarwal, Gaurav. "EnvisionText: Enhancing Text Recognition Accuracy through OCR Extraction and NLP-based Correction." 2024 14th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 2024, pp. 47–51. doi: 10.1109/CONFLUENCE60223.2024.10463478.

[6]. Khan, Z. A.; Xia, Y.; Khaliq, F.; Khan, J. A.; and Khan, N., "From Tradition to Technology: A Systematic Survey on Navigating Pashto in Modern NLP," SSRN, Jan. 2024. Available: https://ssrn.com/abstract=5031721. doi: 10.2139/ssrn.5031721.

[7]. Malashin, I. M.; Masich, I. M.; Tynchenko, V. T.; Gantimurov, A. G.; Nelyub, V. N.; and Borodulin, A. B., "Image text extraction and natural language processing of unstructured data from medical reports," Machine Learning and Knowledge Extraction, vol. 6, pp. 1361–1377, 2024. doi: 10.3390/make6020064.

[8]. S. K. Garai, S. Ghoshal, O. Paul, N. Biswas, U. Dey, and S. Mondal,"A Novel Method for Image to Text Extraction Using Tesseract-OCR,"*American Journal of Electronics & Communication*, vol. 3, no. 2, pp. 8–11, Oct. 2022.

[9]. A. R. Anitha, R. R. Rajeev, M. Nazeem and N. S.,"Open-Source OCR Libraries: A Comprehensive Study for Low Resource Language,"*2024 IEEE International Conference on Networks (ICON)*, Thiruvananthapuram, India, 2024, pp. 1–8,doi: 10.1109/ICON65529.2024.10694684.

[10]. S. Ji, Z. Song, F. Zhong, J. Jia, Z. Wu, Z. Cao, and T. Xu,"OpenGrok: Enhancing SNS Data Processing with Distilled Knowledge and Mask-like Mechanisms,"*arXiv preprint arXiv:2502.07312*, Feb. 2025.

[11]. C. Brogly *et al.*,"Evaluation of the Phi-3-Mini SLM for Identification of Texts Related to Medicine, Health, and Sports Injuries,"*arXiv preprint arXiv:2504.08764*, Apr. 2025.

[12]. Thangam, S.; Kumaran, U.; Biswas, D.; Sneha, B.; Nadipalli, S.; and Raja, S., "Text Post-processing on Optical Character Recognition output using Natural Language Processing Methods," 2023 IEEE 3rd Mysore Sub Section International Conference (MysuruCon), Mysuru, India, 2023, pp. 1–6, doi: 10.1109/MysuruCon59703.2023.10396964.

[13]. Vinusha, B.; Nikhitha, N.; Indhuja, V.; Reddy, V. M.; and Siva Reddy, N. V., "Advancing optical character recognition for handwritten text: Enhancing efficiency and streamlining document management," 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT), 2023. doi: 10.1109/ICCCNT56998.2023.10307143.

[14]. Moussaoui, H.; El Akkad, N.; and Benslimane, M., "License plate text recognition using deep learning, NLP, and image processing techniques," Science of Information and Computing, vol. 1, no. 1, pp. 1–10, 2024. Available: https://www.iapress.org/index.php/soic/article/view/1966.

[15]. Nguyen, T. T. H.; Jatowt, A.; Coustaty, M.; and Doucet, A., "Survey of Post-OCR Processing Approaches," ACM Computing Surveys, vol. 54, no. 6, Article 124, pp. 1–37, July 2021. Available: https://doi.org/10.1145/3453476.

