Automated Medical Labels Detection and Text Extraction Using Tesseract

Imène Ait Abderrahim(✉) and Youcef Benrached(✉)

Khemis Miliana University, Ain Defla, Algeria
{i.aitabderrahim,mi19.y.benrached}@univ-dbkm.dz

Abstract. Medical labels are widely employed in the field of pharmacology to convey critical information about medications, including their names, prices, and other relevant details. In Algeria, the absence of an approved standardized model by the Minister of Health has resulted in the lack of automated identification systems for medical labels. Tesseract has proven to be highly effective in OCR and text extraction tasks, making it a suitable choice for this application. This study introduces an artificial intelligence-based method designed to automatically recognize and extract information from medical labels using Tesseract. Experimental results demonstrate the overall effectiveness of the system, with Tesseract proving particularly effective in text extraction tasks from medical labels images.

Keywords: Image Processing · Text Extraction · Tesseract · Optical Character Recognition (OCR) · LSTM · Medical labels · Healthcare Data Management

1 Introduction

In the contemporary digital age, data is predominantly processed and managed automatically, yet Algerian pharmacies have not fully leveraged this potential regarding medication labels. Currently, these labels are used primarily to identify medication names, while barcodes serve as the primary management tool for selling medicines. The submission of medication information to the National Social Insurance Fund (CNAS) involves attaching medication stickers to the back of prescription sheets, followed by manual verification of the labels against the delivery note (le bordereau). This manual process is time-consuming and prone to errors, prompting the question of how to automate the entire workflow from label detection to data extraction and verification.

Despite significant advancements in optical character recognition (OCR) technology, the recognition of medication labels remains unresolved. This project aims to develop an AI-based OCR system specifically designed for this purpose. By automating the detection and segmentation of information on medication labels, the content can be extracted and stored in a text file, which can then be automatically compared with other documents, such as the delivery note. This

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
S. Chikhi et al. (Eds.): MISC 2024, LNNS 1267, pp. 212–225, 2025.
https://doi.org/10.1007/978-3-031-82112-7_16

automation would streamline the interaction between pharmacies and CNAS, enhancing efficiency and accuracy in medication management. In conclusion, our proposed system leverages AI to fully utilize medication stickers, transforming the current manual process into an automated, efficient, and reliable system. This advancement not only addresses existing challenges but also sets the stage for future improvements in healthcare data management in Algeria.

2 Related Work

Medical labels in several countries provide general information about medications, such as the medicament's name, the number of capsules, the price, and the manufacturing company. However, there is no standardized format for this information. Some labels feature only a barcode, others combine a barcode with the medicament's name, and some display only the medication details. Due to this inconsistency, France opted to eliminate the use of medical labels in 2012, replacing them with barcodes that contain all necessary information[1]. Conversely, Algeria has its own format for medical labels. To enhance the utilization of this information, we propose leveraging advancements in computer vision and image processing to detect and recognize the content of medical labels automatically.

Computer vision operations, such as image identification and segmentation, are fundamental tasks that have recently experienced significant advancements. These tasks involve locating objects within an image, recognizing them, and segmenting them into meaningful sections. Various methods have been proposed for image detection and segmentation, ranging from traditional techniques to advanced deep learning approaches [4]. Traditional methods often involve the application of clustering or segmentation algorithms followed by the extraction of manually crafted features, such as edges, textures, and color information. While these methods have demonstrated reasonable performance, they are limited by their reliance on manually generated features, which may not adequately capture complex patterns [8,11,14,16].

Conversely, deep learning-based methods have emerged as the dominant paradigm for image segmentation and detection. Convolutional Neural Networks (CNNs) have revolutionized artificial intelligence by automatically learning hierarchical features from raw image inputs [9]. Among the prominent deep learning architectures for image identification and segmentation, the YOLO (You Only Look Once) framework stands out, attracting significant attention [12]. YOLO's single-shot detection method is highly efficient for real-time applications, directly predicting bounding boxes and class probabilities by dividing the image into a grid, demonstrating impressive accuracy and speed.

Instance segmentation approaches have also gained popularity for identifying distinct object instances within an image. Mask R-CNN [1], a well-known system, extends the Faster R-CNN object detection architecture with a branch

[1] https://sante.gouv.fr/soins-et-maladies/medicaments/professionnels-de-sante/suppression-de-la-vignette-pharmaceutique-questions-reponses-a-l-attention-des/.

for generating pixel-level segmentation masks, enabling precise instance segmentation crucial for tasks requiring object separation and counting.

Another critical area in computer vision and pattern recognition is text recognition from images, commonly known as optical character recognition (OCR). Text recognition aims to automatically extract and interpret textual information from images, facilitating the conversion of printed or handwritten text into machine-readable formats. This technology has numerous applications, including document analysis, information retrieval, digital archiving, and text-to-speech synthesis.

Text recognition techniques have significantly advanced due to developments in machine learning, deep learning, and image processing. Traditional approaches often combine preprocessing procedures, feature extraction, and classification algorithms. Preprocessing steps like image enhancement, noise reduction, and binarization improve input image quality. Feature extraction techniques, such as connected component analysis or stroke extraction, identify individual text components and their spatial relationships. Classification methods, including template matching or neural networks, are then used to recognize and interpret the extracted text. In recent years, deep learning approaches, particularly CNNs and Recurrent Neural Networks (RNNs) [2,5], have demonstrated remarkable performance in text recognition tasks [15].

Another prominent area of study in computer vision and pattern recognition is text recognition from photographs, sometimes referred to as optical character recognition (OCR). To enable the conversion of printed or handwritten text into machine-readable formats, text recognition aims to automatically extract and interpret textual information inherent within photographs. Document analysis, information retrieval, digital archiving, and text-to-speech synthesis are just a few of the many uses of this technology.

Techniques for text recognition have also advanced significantly on their turn as a result of developments in machine learning, deep learning, and image processing. Traditional approaches to text recognition frequently combined feature extraction, preprocessing procedures, and classification algorithms. To enhance the quality of the input image, preprocessing processes including image enhancement, noise reduction, and binarization are used. Methods for identifying individual text parts and their spatial relationships, such as linked component analysis or stroke extraction, are called feature extraction techniques. To detect and analyze the retrieved text, classification methods like template matching or neural networks are used. on the other hand, in recent years, deep learning approaches [2], particularly Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) [5], have shown remarkable performance in text recognition tasks [15].

In addition to these advanced algorithms, Tesseract-based methods remain important in the field of OCR. Tesseract, an open-source OCR engine first created by Hewlett-Packard and later upgraded by Google, takes a classical approach to text recognition. It uses adaptive thresholding, connected component analysis, and pattern matching to detect and recognize characters[13]. To detect

and recognize characters, Tesseract uses a multistage approach that includes adaptive thresholding, linked component analysis, and pattern matching. Recent developments have integrated LSTM (Long Short-Term Memory) networks to handle line recognition, making it stronger for diverse text recognition tasks [13]. Tesseract has been shown to be effective in a variety of applications, including identifying text in historical texts and extracting information [6,7]. Tesseract's inclusion into broader image processing processes remains valuable, particularly when combined with standard techniques for preprocessing and segmentation [3]. This hybrid method guarantees that Tesseract remains a robust and adaptable tool for OCR jobs, supplementing the more complex deep learning-based systems.

3 Tesseract-Based Approach

In this section, we describe our OCR-based approach using the Tesseract from Meghatria et al. [10].

3.1 Dataset Collection and Organization

Given the absence of an Algerian Medical Labels dataset, we compiled an image dataset by photographing labels from a local pharmacy called "Dis Mebarek" in Khemis Miliana, Ain Defla, Algeria. This dataset consists of over 1,000 images of 300 different medications. The labels were photographed from various angles and positions, as illustrated in Fig. 1.

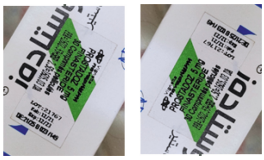

Fig. 1. Medical label Image

Following data collection, the labels were organized into a tabular format to facilitate later exploration as shown in Fig. 2). This table includes details such as the image ID, the number of textual sections on the medical label, its color (red, green, or white), and the text content of each section.

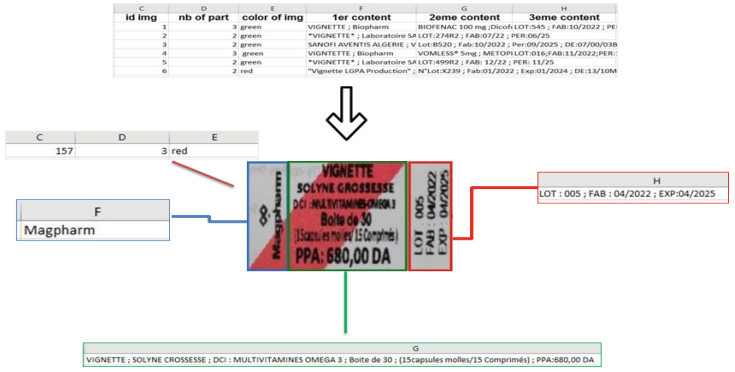

Fig. 2. Medical labels tabular documentation

3.2 Tesseract-Based Approach

The Tesseract-based method begins with the manual identification of the four corners of the medical label and proceeds using a traditional semi-intelligent approach. Histogram analysis is subsequently employed to detect the lines of text within these sections, as well as their vertical and horizontal components. Character recognition algorithms are then utilized to extract meaningful information.

This technique leverages conventional detection and segmentation methods, such as histogram thresholding for image segmentation. The aim is to achieve precise and reliable identification and segmentation of medical labels. The identified segments are subsequently passed to the Tesseract recognition module, which processes each line separately and ultimately compiles the lines to generate the recognized text from the image.

The following flowchart in Fig. 3 illustrates the steps involved in text recognition using the Tesseract-based approach.

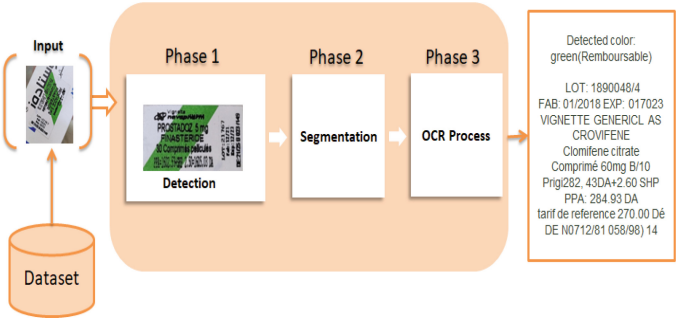

Fig. 3. Architecture of the Tesseract-based approach

Phase 1: Medical Label Detection

This phase constitutes the initial step of the Tesseract-based method. It involves manually selecting the four corners of the medical label by clicking with a mouse. After completing the selection, the system applies a corrective rotation to properly align the label.

Fig. 4. Phase1: Detection in Tesseract-based approach

Phase 2: Segmentation Phase

Medical labels contain a horizontal section with details like the manufacturer's name, medication name, dosage, and price, and a vertical section with the lot number, expiration date, and manufacturing date. To analyze and differentiate these sections, the image is converted to gray-scale. The gray-scale histogram is used to identify gaps, allowing for the segmentation of the image. Vertical sections are rotated counterclockwise for alignment, and text lines within these sections are separated using a horizontal histogram. The positions of the remaining segments are recorded and applied to the original color image for accurate segmentation. This process is repeated for each part of the label.

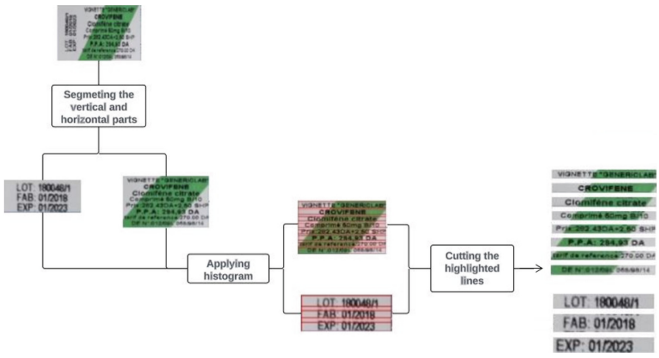

Fig. 5. Phase2: Segmentation in Tesseract-based approach

Phase 3: Character Recognition

In this stage, each line of text undergoes immediate optical character recognition (OCR) using the PyTesseract OCR model individually.

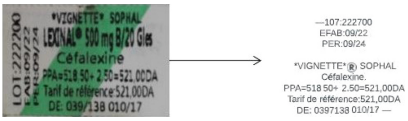

Fig. 6. Tesseract approach output

4 Proposed Fine-Tuned Tesseract Approach

One of the issues that the Tesseract approach in [10] encounters is that it does not correctly detect the text within the medical labels due to their bad printing quality and the lack of a standard model. In this section, we propose fine-tuning the model from [10] using a separate dataset that we manually created for this purpose of tuning. To do so, the fine-tuning process is mainly composed of two phases, the first phase consists of the data creation and the second phase refers to the fine-tuning process. Furthermore, to evaluate the quality of the model's performance where two metrics have been considered: *Accuracy* and *loss* (BCER). All the steps are described in more detail as follows:

4.1 Dataset Creation

Fine-tuning the data is a crucial step to enhance the performance quality of the Tesseract approach. In the context of Tesseract, combining text images, ground truth, and box files into a Long-Short Term Model (LSTM) file for fine-tuning refers to a specific process in preparing and optimizing the OCR model for better performance.

Data Preparation

In this process and as shown in Fig. 7, we first generate a dataset comprising images of text in *.tif* format along with their corresponding ground truth text files. This involves extracting text lines from the medical labels, creating text files containing the actual text for each line, and ensuring that each image and its associated text file share the same name. Then, in the next step, we would be creating the box files for each of these images using the ground truth text files which serve as the correct reference for the OCR system to learn from, see Fig. 8.

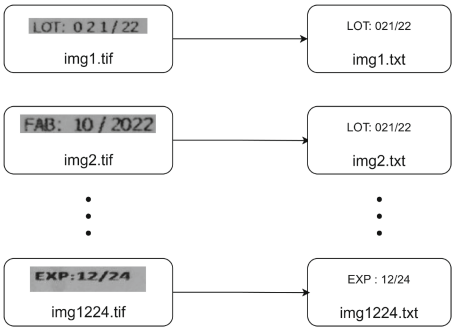

Fig. 7. Dataset preparation illustration example

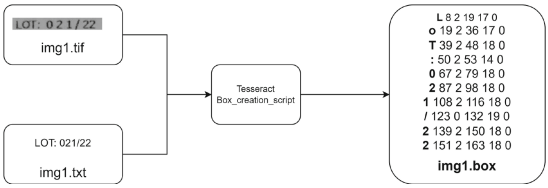

Fig. 8. Box File Creation

Training Data Creation

In this step, the text images, ground truth, and box files are combined into a training dataset that Tesseract's training process can use. This dataset is used to train the LSTM model on specific types of text images.

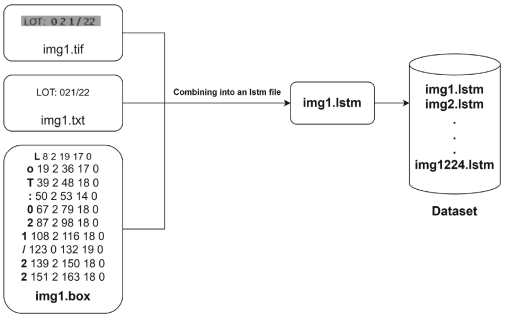

Fig. 9. LSTM File Creation

4.2 Fine Tuning Process

During fine-tuning, the existing OCR model is further trained on this prepared dataset. The process begins by extracting the pre-trained base LSTM neural network from the standard Tesseract OCR engine. This extracted LSTM serves as the foundation for our training process. Following the extraction, the model is provided with the *.traineddata* file, which contains the essential trained data required for the OCR engine to recognize text. This includes the neural network parameters, character set, and language-specific data. The *.traineddata* file can be accessed via the official Tesseract repository on GitHub[2]. In addition to the *.traineddata* file, we specified our custom dataset with the LSTM files, as seen in Fig. 10. This custom dataset is used to train and fine-tune the basic LSTM model.

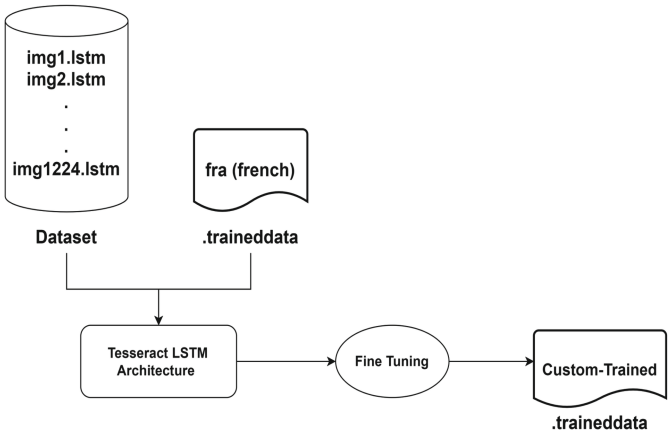

Fig. 10. Tesseract Fine Tune: Fine Tuning Process

4.3 Fine-Tuning Evaluation

The presented graphs illustrate the progress of the Tesseract OCR fine-tuning.

Training Accuracy Evaluation

The graphic illustrated in Fig. 11 tracks the model's accuracy improvement over 350 iterations. The x-axis represents the number of iterations, while the y-axis displays accuracy as a percentage. The graphic shows a steady increase in accuracy from approximately 80% to 85%. This upward trend indicates that the model is progressively learning and improving its ability to accurately recognize and interpret medical labels.

[2] https://github.com/tesseract-ocr/tessdata_best/blob/main/fra.traineddata.

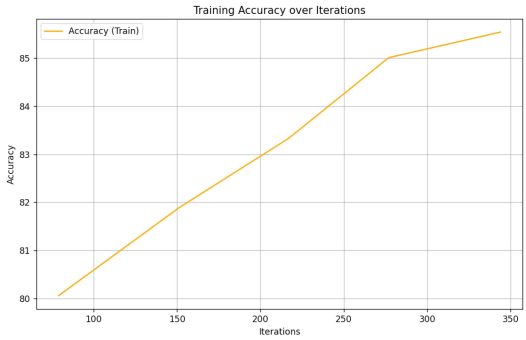

Fig. 11. Accuracy Graph during Fine-Tuning

Training Loss (BCER) Evaluation

The graphic in Fig. 12 illustrates the reduction in the model's training loss over a series of iterations. The x-axis represents the number of iterations, while the y-axis displays the Binary Cross-Entropy Loss (BCER). BCER is a common loss function for binary classification problems, measuring the performance of a classification model whose output is a probability between 0 and 1. It shows a consistent decrease in BCER from approximately 20 to 15. This downward trend in training loss indicates that the model's predictions are becoming more accurate and better aligned with the actual labels, reflecting effective learning from the training data.

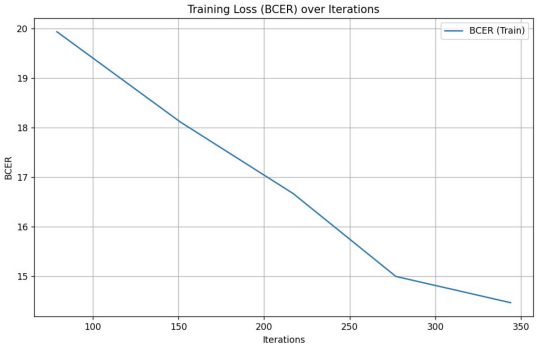

Fig. 12. Accuracy Graph during Fine-Tuning

Upon completion of the fine-tuning process, the updated Tesseract *.trained-data* file, now better adapted to the specific characteristics of the text in the medical labels, is saved. This refined model is then compiled into a new *.trained-data* file, which integrates the original base model data with the newly acquired features. The resulting *.traineddata* file is now ready for use by Tesseract for

OCR tasks, offering enhanced accuracy and reliability tailored to the specific use case for which it was fine-tuned.

5 Experimental Results

Preliminary tests on 40 images using the standard Tesseract OCR model, without fine-tuning, showed promising performance overall. Despite this, some letters were misrecognized, particularly those with distinctive features like the initial letters in medication names, which are often decorated with large capital letters. Additionally, characters with poor printing quality posed recognition challenges. These minor discrepancies provide valuable insights for future refinement and optimization of the model. Therefore, Fine-tuning the standard Tesseract OCR model was necessary to improve its character recognition accuracy in contexts where it initially struggled. By training the model on a dataset specifically targeting issues such as decorated large capital letters and poor printing quality, recognition accuracy can be enhanced. The fine-tuned model aims to better accommodate these unique text characteristics and handle variability in text presentation more effectively. This process is expected to reduce misrecognition rates and enhance overall performance, particularly in scenarios where the standard model was insufficient.

For comparison, figure Fig. 13 illustrates the recognized texts of the same medical label using each of the Tesseract models, standard and fine-tuned, respectively. This figure shows some visual contrast and assessment of the performance of each OCR engine.

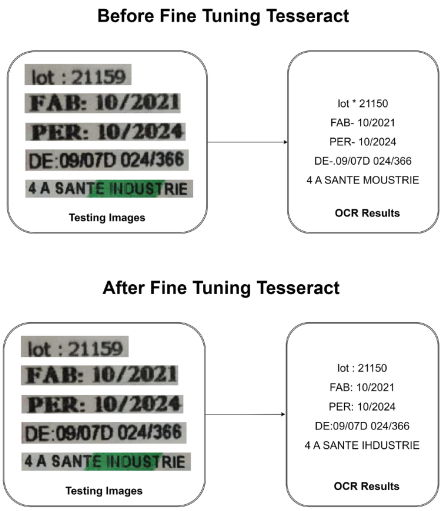

Fig. 13. Tesseract results comparison before and after the fine tuning

Additional experiments were conducted on forty new medical label images, distinct from those in the training and validation sets, to assess the effectiveness of the proposed approach using the standard model. The accuracy and reliability of the text identification were evaluated by comparing the recognized text to the original text on the medical labels, utilizing the following formula (1):

$$accuracy = \frac{NCCR \times 100}{TNC} \tag{1}$$

where NCCR stands for "Number of Characters Correctly Recognized" and TNC stands for "Total Number of Characters in the medical label".

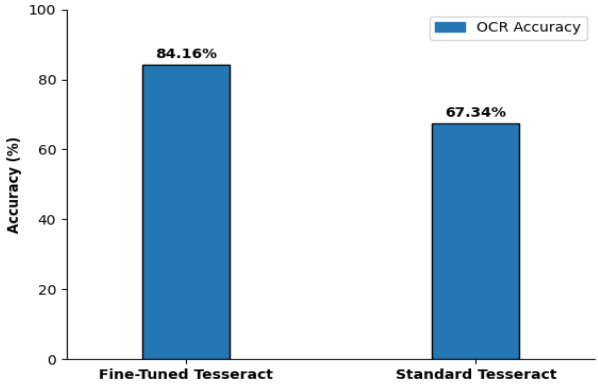

Fig. 14. Comparison study results

As shown in Fig. 14, the standard Tesseract OCR achieved a text recognition accuracy of 67.34% on 40 medical labels, which, while acceptable, suggests a need for further improvement. In contrast, the fine-tuned Tesseract model attained a significantly higher accuracy of 84.16% on the test set. This improvement highlights that fine-tuning enhances the OCR model's ability to accurately recognize text from medical labels and adapt to specific text characteristics. The results underscore the value of customized training for achieving greater accuracy and reliability in real-world applications.

These performance outcomes from the standard Tesseract model can be attributed to several factors. Primarily, the use of the Standard Tesseract OCR model in its default configuration, without any fine-tuning or additional training, impacted the results. The Standard Tesseract version 4.0 employs a neural network for text detection but lacks specialized customizations beyond its general training.

Conversely, the superior performance of the Fine-Tuned Tesseract is due to its targeted training on a specialized dataset designed to address the specific challenges of recognizing medical labels. This fine-tuning process customizes the OCR model to better handle unique text characteristics and contexts, leading

to significantly improved accuracy. The enhanced results highlight the advantages of customizing OCR models for specific text recognition tasks, providing more reliable and precise outcomes compared to the standard model.

6 Conclusion

This study presents an AI-based system for text extraction from medical labels in Algeria, addressing the gap in standardized recognition systems. By evaluating a Tesseract-based approach and a fine-tuned variant of the Tesseract approach solutions, our experiments reveal that the fine-tuned Tesseract model achieved a notable accuracy of 84.16% in label detection and content extraction. This performance improvement underscores the benefits of customizing OCR models for specific text characteristics. The fine-tuned Tesseract model offers significant advantages for pharmacies, insurance providers, and CNAS by enhancing medication handling and streamlining procedures. The research demonstrates the potential of AI technologies to improve healthcare outcomes nationwide. Future work will focus on further fine-tuning Tesseract and exploring additional AI approaches to advance medical label detection and text extraction.

References

1. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2961–2969 (2017)
2. Hossain, M.Z., Sohel, F., Shiratuddin, M.F., Laga, H.: A comprehensive survey of deep learning for image captioning. ACM Comput. Surv. (CsUR) **51**(6), 1–36 (2019)
3. Kaundilya, C., Chawla, D., Chopra, Y.: Automated text extraction from images using ocr system. In: 2019 6th International Conference on Computing for Sustainable Global Development (INDIACom), pp. 145–150. IEEE (2019)
4. Khan, W.: Image segmentation techniques: a survey. J. Image Graph. **1**(4), 166–170 (2013)
5. Koutnik, J., Greff, K., Gomez, F., Schmidhuber, J.: A clockwork rnn. In: International Conference on Machine Learning, pp. 1863–1871. PMLR (2014)
6. Kumar, S., Sharma, N.K., Sharma, M., Agrawal, N.: Text extraction from images using tesseract. Deep Learn. Tech. Autom. Ind. Appl. 1–18 (2024)
7. Kumar Garai, S., Paul, O., Dey, U., Ghoshal, S., Biswas, N., Mondal, S.: A novel method for image to text extraction using tesseract-ocr. Am. J. Electron. Commun. **3**(2), 8–11 (2022)
8. Lakshmi, S., Sankaranarayanan, D.V., et al.: A study of edge detection techniques for segmentation computing approaches. IJCA Special Issue on "Computer Aided Soft Computing Techniques for Imaging and Biomedical Applications" CASCT, pp. 35–40 (2010)
9. LeCun, Y., et al.: Handwritten digit recognition with a back-propagation network. Adv. Neural Inf. Process. Syst. **2** (1989)
10. Meghatria, R., Ait Abderrahim, I.: Detection and recognition approaches for medical labels in Algeria. In: 2024 8th International Conference on Image and Signal Processing and their Applications (ISPA), pp. 1–6 (2024). https://doi.org/10.1109/ISPA59904.2024.10536801

11. Nadernejad, E., Sharifzadeh, S., Hassanpour, H.: Edge detection techniques: evaluations and comparisons. Appl. Math. Sci. **2**(31), 1507–1520 (2008)
12. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 779–788 (2016)
13. Smith, R.: An overview of the tesseract ocr engine. In: Ninth International Conference on Document Analysis and Recognition (ICDAR 2007), vol. 2, pp. 629–633. IEEE (2007)
14. Succi, A., Torre, V.: A new approach to image segmentation. In: Image Analysis and Processing: 8th International Conference, ICIAP'95 San Remo, Italy, September 13–15, 1995 Proceedings 8, pp. 17–22. Springer (1995)
15. Ye, Q., Doermann, D.: Text detection and recognition in imagery: a survey. IEEE Trans. Pattern Anal. Mach. Intell. **37**(7), 1480–1500 (2014)
16. Zaitoun, N.M., Aqel, M.J.: Survey on image segmentation techniques. Procedia Comput. Sci. **65**, 797–806 (2015)