SSRG International Journal of Computer Science and Engineering
ISSN: 2348–8387 / https://doi.org/10.14445/23488387/IJCSE-V11I11P104

Volume 11 Issue 11, 25-34, November 2024
© 2024 Seventh Sense Research Group®

Original Article

Intelligent Data Extraction from Image Documents

Dhivya Nagasubramanian

Lead AI Solutions Architect, AI Process Transformation and Automation.

Corresponding Author : nagas021@alumni.umn.edu

Received: 29 September 2024 Revised: 28 October 2024 Accepted: 13 November 2024 Published: 30 November 2024

Abstract *- Enterprises often possess a vast collection of scanned documents and images with valuable data crucial for organizational growth and success. In the finance industry, for instance, banks manage extensive collateral documents, tax forms, title deeds, and other critical materials, such as check images, syndication records, and flood documentation. Extracting information from these extensive, scanned files typically involves manual data entry, which is time-consuming and susceptible to human error. With advancements in AI, document entity extraction can now be automated in multiple ways. Heuristic methods can be employed for simpler documents where entities consistently appear in predefined spaces. More complex scenarios can leverage AI frameworks, such as Convolutional Neural Networks (CNNs), trained on labeled images to detect regions of interest, producing bounding boxes and confidence scores for the predictions. Generative AI toolkits offer another solution: extracting entities directly from documents or facilitating question-and-answer interactions to retrieve specific information efficiently. This research paper explores how these methodologies can be swiftly adopted based on document complexity, evaluates the advantages and limitations of each approach, and discusses the role of pipeline building in enhancing the accuracy of AI model predictions.*

Keywords *- Document intelligence, Document extraction, CNN, Convolutional neural network, Transformer, OCR, Optical character recognition, Encoder-decoder model, Object detection, Layout detection.*

1. Introduction

The financial industry handles many documents, many of which are intricate and contain jargon unique to the sector. These documents can have different formats and structures depending on where they came from. Internally developed collateral documents, for example, might follow conventional templates, whereas externally prepared collateral documents might have completely distinct formats. Regional differences in loan terms and agreements make the work even more difficult.

It needs sophisticated tools to decipher complex language to extract important information like maturity dates, interest rates, floor percentages, agreement dates, and borrower characteristics. Contemporary AI technologies are frequently created for specific tasks, such as text generation, image-to-text conversion, and object identification. This study examines how various technologies might be combined and tailored to meet the specific needs of processing financial documents. For simpler documents with predictable structures, such as money orders, checks, tax forms, and financial statements, a combination of heuristic methods and Convolutional Neural Networks (CNNs) works well. CNNs, a type of deep learning framework, are very good at identifying text in scanned documents since they are made to interpret structured data, such as things in an image. Because the printed text is homogeneous, these models perform remarkably well when identifying and isolating sections of interest by drawing bounding boxes, frequently with high confidence levels. Nevertheless, they become less successful when handling handwritten text, which varies greatly in terms of shape and style. Advanced models such as transformers are essential for more unstructured and complex documents. Transformers have transformed natural language processing through their ability to analyze and interpret context within lengthy texts. Even when training data is scarce, these models perform exceptionally well on tasks such as named entity recognition. Unstructured document analysis is a good fit for them because they can also handle handwritten information more efficiently with less input than traditional models. In order to close the gaps in earlier research, this study investigates methods for automating data extraction from unstructured documents. The effectiveness of various approaches—from simple rule-based systems to cutting-edge transformer models—is evaluated according to how well they manage complex documents, extract crucial data and function effectively in practical situations. The evaluation uses metrics such as accuracy, speed, and overall efficacy. The dataset was split into 70% for training and 30% for testing to ensure fairness in comparisons. All models were tested on the same document set to maintain consistency. Understanding that a single solution cannot address all scenarios, this paper explores multiple strategies. It offers guidance on selecting the most appropriate techniques based on document complexity and specific requirements.

2. Related Work

[1] Vaswani A. et al. introduced Transformer architecture, a groundbreaking sequence transduction model that is purely based on attention mechanisms. By eliminating the need for recurrence and convolutions, the Transformer achieves greater performance in machine translation, setting new benchmarks for BLEU scores. Its impact extends beyond machine translation, with subsequent success in tasks like English constituency parsing. The Transformer's ability to scale effectively to large datasets and its generalizability to various NLP tasks make it a vital component in modern NLP pipelines. Building on this foundation, the proposed work leverages transformer-based models like LayoutLMv3 and TrOCR, fine-tuned for financial document processing, demonstrating superior results in extracting entities from complex and unstructured data. [2] Wang, C.-Y., et al. present YOLOv7, a real-time object detection model that significantly improves both speed and accuracy compared to previous versions of YOLO, as well as other state-of-the-art detectors like SWIN-L Cascade-Mask R-CNN. YOLOv7 achieves an impressive 56.8% average precision (AP) at 30 FPS or higher, surpassing prior models by over 500% speed. Trained on the MS COCO dataset, YOLOv7 achieves higher accuracy and faster inference times than earlier YOLO models, making it an excellent choice for real-time applications such as signature detection and text entity extraction in financial documents. In this research, YOLOv7's real-time detection capabilities were adapted to detect critical entities in scanned financial documents, demonstrating its efficiency in identifying signatures and bounding boxes.

[3] Liu, W. et al. introduced the SSD (Single Shot MultiBox Detector), a model designed to efficiently handle object detection at different scales without requiring proposal generation or pixel resampling. SSD outperforms Faster R-CNN in both speed and accuracy, achieving a mean Average Precision (mAP) of 72.1% at 58 FPS on 300×300 input and 75.1% mAP on 500×500 input. Unlike YOLOv7, which uses a single, unified architecture for object detection, SSD uses multiple feature maps to discretize bounding boxes across various aspect ratios and scales. This multi-scale approach makes SSD particularly effective for detecting objects of varying sizes and shapes. It is suitable for applications such as financial document processing, where object size variation is common. This research complements SSD's strengths by applying it to detect multi-scale objects like MICR lines and text regions in financial documents. Slide Gestalt introduces an innovative approach to improving accessibility for Blind and Visually Impaired (BVI) users by detecting hierarchical patterns in presentation slides. This system dynamically structures slide decks, allowing users to navigate them efficiently. Evaluations on a large dataset of slide decks showed that Slide Gestalt improves content navigation, highlighting its potential for improving accessibility in presentation tools. Although not directly related to document processing, the methodology of hierarchical structuring in Slide Gestalt could inspire similar techniques for structured document analysis, especially in cases where documents require dynamic restructuring for better accessibility. This work borrows the idea of hierarchical structuring for financial documents, using it to enhance entity localization and extraction accuracy. LayoutLMv3, a multimodal Transformer that combines text and image data for better content comprehension, is presented by Huang, Y. et al. Tasks involving document layout analysis, such as comprehending forms and receipts and responding to inquiries based on visual context, are especially well-suited for this paradigm. By utilizing both textual and visual clues, LayoutLMv3 surpasses previous document analysis models and provides a cohesive method for multimodal document understanding. For financial document processing, LayoutLMv3 provides a strong solution for hierarchical understanding of complicated documents, where layout and text frequently transmit important information. This study fine-tuned LayoutLMv3 to excel in extracting structured and unstructured data from checks, leveraging its multimodal capabilities for accuracy improvements. This research extended NER principles using Transformer-based models to handle financial documents with multi-language support.

[4] Lample, G. et al. explore neural architectures for Named Entity Recognition (NER), combining bidirectional LSTMs with Conditional Random Fields (CRFs) for improved sequence labeling. Their models achieve state-of-the-art NER performance in four languages without relying on domain-specific knowledge or gazetteers. Using unsupervised word embeddings and character-based representations improves accuracy, making the model versatile across multiple languages. This work underscores the effectiveness of neural networks for NER, which is essential for applications like document text extraction and information retrieval in financial datasets. This research extended NER principles using Transformer-based models to handle financial documents with multi-language support. [5] He, K. et al. introduce Residual Networks (ResNets), which use residual learning to facilitate the training of very deep neural networks. ResNets overcome optimization challenges in deep networks, demonstrating significant improvements in performance on tasks like image classification and object detection. The ResNet framework has become foundational in modern computer vision models, and its principles are applied in many state-of-the-art object detection models, such as YOLOv7 and SSD. These techniques can be adapted for deep learning systems that recognize complex visual patterns in documents, such as signatures or handwritten text. This study applies ResNet's architecture in feature extraction stages, enhancing document entity detection accuracy across both CNN and Transformer models. [6] Rajpurkar, P. et al. present the Stanford Question Answering Dataset (SQuAD), a benchmark for reading comprehension tasks. While the authors' logistic regression model significantly outperforms the baseline, it still falls short of human-level performance. SQuAD has driven

advancements in machine comprehension and remains a key dataset for training and evaluating models to improve document understanding, especially in scenarios where extracting specific answers from a passage is required. This work informs approaches in Document AI, particularly in fine-tuning models for question-answering tasks in document processing. This work adapts similar comprehension-based approaches to extract financial document entities, addressing both structured and unstructured data contexts.

[7] Ren S. et al. introduce the Region Proposal Network (RPN), which significantly improves object detection by generating region proposals that share convolutional features with the detection network. RPN enables efficient object detection through end-to-end training, facilitating the integration of attention mechanisms for better focus on relevant regions. This method has been crucial in systems that require precise object localization, such as document analysis systems that detect and classify regions of interest, including signatures or text entities. This work incorporates region proposals into document layouts for more accurate localization of entities like signatures and MICR lines. [8] Redmon, J. et al. present YOLO (You Only Look Once), a real-time object detection system that simplifies detection by treating it as a regression problem. YOLO processes entire images at once, enabling high-speed detection with lower computational costs. While the original YOLO achieves 45 FPS, the Fast YOLO version further optimizes speed, processing images at 155 FPS. YOLO's approach has set a new standard for real-time detection, making it a valuable tool in real-time document analysis applications. This research builds on the YOLO architecture by fine-tuning YOLOv7 to handle domain-specific complexities in financial documents. Lewis R. et al. highlight the challenges in evaluating OCR and document analysis technologies due to the lack of publicly available, realistic test sets. Their 1.5-terabyte dataset provides an extensive resource for evaluating technologies like OCR and signature matching, which are crucial for improving document processing pipelines. This study addresses these challenges by creating annotated datasets tailored to financial documents for robust model evaluation. [9] Smith, R. reviews the Tesseract OCR engine, known for its robust and versatile capabilities in document image processing. Tesseract's line-finding algorithm, feature extraction methods, and adaptive classifier set it apart as a powerful OCR tool for various document processing applications. Despite being a robust solution, Tesseract is limited by its performance in complex document layouts, making newer systems like TrOCR or LayoutLMv3 more suitable for multimodal document tasks. The current study addresses Tesseract's limitations by integrating TrOCR for higher accuracy in handwritten and multi-format document processing. Li H. et al. introduce TrOCR, a Transformer-based OCR system that utilizes pre-trained models to perform text recognition on printed and handwritten documents. TrOCR's high adaptability and state-of-the-art performance across OCR tasks demonstrate its superiority over older OCR models like Tesseract. The Transformer architecture in TrOCR allows for better handling of varied document types, especially in scenarios where high accuracy is critical. This research optimizes TrOCR's performance for handwritten financial documents, achieving significant improvements in accuracy and reliability. [10] FLAIR, a unified framework for NLP tasks, is presented by Akbik, A. et al. and makes training models for tasks like text categorization and sequence labeling easier. FLAIR simplifies NLP development by abstracting embedding-specific difficulties, making it simple for researchers to test various models. This framework is useful for quickly prototyping and deploying NLP models, especially when incorporating sophisticated models into document-based applications. This work incorporates FLAIR for pre-labeling text in financial documents, improving overall model precision and consistency. [11] Lafferty, J. et al. discuss Conditional Random Fields (CRFs), a probabilistic framework designed to improve sequence labeling tasks by relaxing independence assumptions. Their work highlights the effectiveness of CRFs in handling complex sequence data, which has applications in tasks like NER and document information extraction, where the data structure plays a critical role. This research replaces CRFs with Transformer-based contextual embeddings for superior performance in recognizing entities from unstructured text. Vedhaviyassh et al. propose a multi-module framework for license plate recognition, using YOLOv5 for detection and EasyOCR for character recognition. By integrating the strengths of object detection and OCR, the framework achieves high accuracy and speed. While designed for license plate recognition, this methodology applies to document analysis systems, where high-speed character and entity extraction is necessary. This research adapts a similar multi-module design for detecting and extracting entities from checks using YOLOv7 and TrOCR. This research adapts a similar multi-module design for detecting and extracting entities from checks using YOLOv7 and TrOCR.

Tan, Y. et al. present the Dynamic Weighting Structural Analysis (DWSA) framework for hierarchical classification and structural analysis of documents, particularly focusing on business contracts and legal documents. The framework incorporates data preprocessing, feature engineering, structural classification, and a dynamic sample weighting algorithm to address imbalanced data. With a comprehensive accuracy improvement of 94.68% and a Macro F1-score of 88.29%, the DWSA model significantly outperforms baseline approaches in hierarchical structural analysis. This research applies dynamic weighting to prioritize critical entities in financial documents, enabling more accurate recognition of complex layouts. [12] The paper reviews information extraction techniques applied to medical documents, focusing on Named Entity Recognition (NER) and relation extraction. It discusses the application of Natural Language Processing (NLP) methods for extracting critical information such as drugs, diseases, and patient details from unstructured narratives. The study outlines potential future research directions in the medical domain by addressing challenges like semantic ambiguity and the large volume of medical records. This research adapts these techniques for extracting structured entities such as payer and payee information from financial documents. [13] Doermann D. et al. provide an updated survey of research in document-based information retrieval, emphasizing the connection between document image analysis methods and traditional IR techniques. The study covers methods for handling text and image documents, including OCR and direct image manipulation. It also discusses challenges like noisy text from OCR and content-based image retrieval. This paper bridges these IR methods with modern AI techniques, enabling more efficient and accurate financial document analysis.

3. Methodology

Check images typically have a straightforward layout, with each entity appearing in a predefined space. Entities that can be extracted from check images include payee name, payer name, payee address, payer address, check amount, amount in words, date of issue, issuing bank, signature, and MICR number. For example, the MICR number is always at the bottom, the payer's name and address are at the top, and the check amount in numbers is on the right. Extracting these entities becomes significantly more complex when dealing with handwritten elements. Heuristic methods for entity extraction are generally more effective on printed documents than handwritten ones. This complexity led to experimenting with multiple models to address the same problem, ultimately selecting the best model for each specific entity. For instance, the confidence or accuracy scores for each entity are compared across three different models, and the model with the highest score for a given document-entity combination is chosen as the final winner.

This research paper explores three approaches for data extraction from images:

1. Heuristic entity extraction method for simple documents (e.g. Check images)
2. Signature and text entity detection (Using YOLOv7 and SSD architecture for object detection)
3. Transformer-based entity extraction.

3.1. Heuristic Method Using OCR

In heuristic modeling, the check was divided into five different regions with some degree of overlap: Quadrant 1 (top left of the check), Quadrant 2 (top right of the check), Quadrant 3 (bottom left of the check), Quadrant 4 (bottom right of the check), and the mid-section spanning from left to right. After splitting the check into these quadrants and the mid-section, OCR on each region was performed using EasyOCR. It is a lightweight OCR model that supports multiple languages and uses a deep learning framework as its backbone.

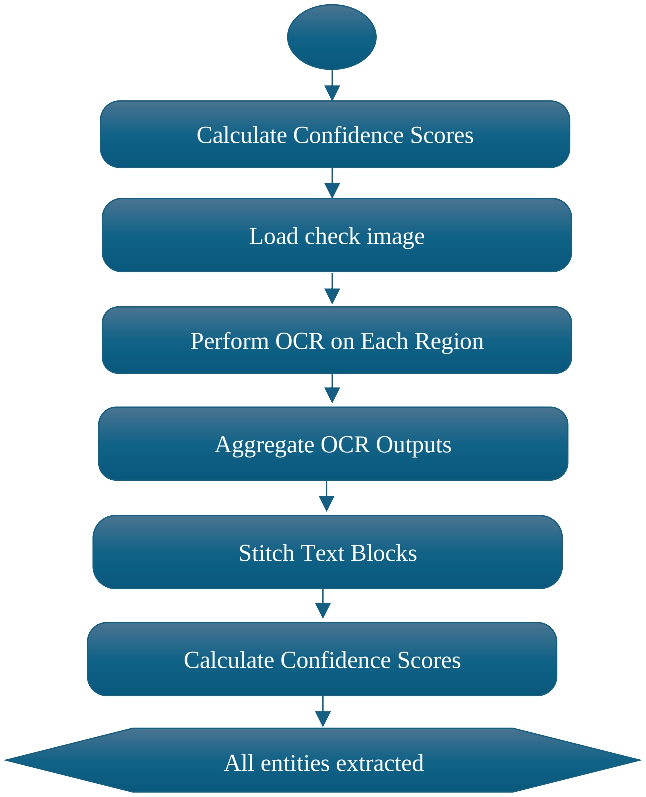
Calculate Confidence Scores
Load check image
Perform OCR on Each Region
Aggregate OCR Outputs
Stitch Text Blocks
Calculate Confidence Scores
All entities extracted

Fig. 1 Process flow diagram for Heuristic framework

Its prediction accuracy is relatively higher compared to other OCR frameworks. The OCR model outputs words, and their relative position coordinates on the check. A heuristic model was then built around the OCR results, which traverses from the extreme top and left towards the right and bottom to stitch similar text blocks and group them as one entity. The confidence scores of these individual text blocks or chunks are averaged to determine the confidence score of the entire entity block. However, this model could not extract signatures from the checks and was not explicitly included in metric comparisons for the signature entity.Heuristic models are particularly applicable, reliable, and faster when the document or image follows a predefined layout. The metrics from this model are compared to those from other models developed for this problem to establish a benchmark model performance for various entities in the check image.

3.2. Signature Detection and Text Entity Detection

Documents like checks are valid only when signed and dated, making signature detection essential for ensuring the text's validity. The developed signature detection solution uses cascaded models for extracting signatures. We use a one-stage model called YOLOv7 (You Only Look Once) for signature detection. In addition, a one-stage model called the Single Shot MultiBox Detector (SSD) model was employed for object detection. The object detection model can detect multiple classes, including signatures, dates, addresses, amounts, etc. It was observed that the model required a significant amount of training examples to perform well, particularly on signature detection. The Tobacco-800 dataset and sample check documents were used to train the YOLOv7 and SSD models effectively. It is worth noting that image scans often come with varying gradients of quality, making it critical to standardize these images to consistent quality and format before tagging elements for training. Models like YOLOv7, for instance, require specific image sizes (such as 640 x 640 pixels), and standardizing them is essential for improving model performance. For image standardization, careful attention was given to resizing images to an optimal size that improved model performance. For example, images were resized to 300 x 300 pixels for the SSD model and 640 x 640 pixels for YOLOv7. Image size is crucial to both model performance and inference speed. The pixel values were normalized to enhance model convergence. Additionally, various data augmentations were applied, including rotation, scaling, cropping, contrast adjustments, adding noise, and applying Gaussian blur, among others, to improve the robustness of the model.

In this case, YOLOv7 and SSD are object identification models trained using this preprocessed input. A neural network recognizes items in a picture and highlights them with bounding boxes and a confidence score in object detection, a sophisticated type of image classification. An essential component of this study was to track each entity's performance results during training. The accuracy and speed of object detection are frequently assessed using standard metrics like IOU (Intersection over Union), recall, precision, F1 score (F1), Mean Average Precision (MAP), etc. Recall is the ratio of correctly identified intended results to all real values in the sample set. In contrast, precision is the ratio of correctly predicted values to all predicted values in the entity detection process. The model's overall performance is evaluated using MAP, which is the average value of the sum of APs (Average Precision) for all entities. Each unique entity's MAP is determined using a particular methodology or calculation.

Example: The output of a YOLOv7 or SSD model consists of bounding boxes and confidence scores for each class the model is trained on. For an entity like "Signature," the evaluation process is as follows:

1. Sort predictions by confidence scores.
2. Determine if each prediction is a True Positive (TP) or a False Positive (FP) based on the IoU (Intersection over Union) score with the ground truth. The user sets the IoU threshold.
3. Compute precision and recall at various confidence score thresholds.
4. Calculate the Average Precision (AP), which is the area under the precision-recall (PR) curve.
5. Repeat this process for all entities and then compute the mean of all APs to get the mean Average Precision (mAP).

It was interpretable from the results that both SSD and YOLOv7 outperformed the heuristic models for other elements in the check image and were equivalent for signature detection. The entity type (date, signature, amount, address, etc.) and the bounding box coordinates are included in the output of YOLOv7. The task was more difficult since the text inside these bounding frames might be handwritten or printed.

For printed text, OCR techniques like EasyOCR, combined with the previously discussed heuristic methodology, were used to group blocks of text and derive the entity value. Resource-intensive language models such as Vision Transformers and their variants to extract handwritten text from images were adopted for handwritten text. In the case of check images, token size limits were not a concern. Iterating through the list of entities and their corresponding bounding boxes from the CNN models, passing the images through the TrOCR model, which is specifically trained on handwritten documents, for inference. This approach helped achieve high accuracy rates, exceeding 95%, even for handwritten text.

3.3. Transformer-Based Entity Extraction

Although checks generally follow a structured layout, they can become highly complex during inference, especially when dealing with handwritten text, combined handwritten and printed text, poor image quality, or excessive background noise. In such cases, transformer models have demonstrated the ability to achieve human-level accuracy.

A cascade modeling solution is employed using LayoutLMv3 for object detection and TrOCR for text extraction (handwritten and printed). LayoutLMv3 is a self-supervised transformer model developed by Microsoft Corporation specifically for document understanding, and it has proven effective in handling the complexities mentioned above. Over 1,000 check images were labeled using an open-source LabelMe tool and saved in JSON format. The labeled dataset, consisting of these 1,000+ documents, was used to fine-tune the LayoutLMv3 model, adapting it to the new business problem of check entity extraction.

Fine-tuning this transformer model is resource-intensive, requiring GPU-enabled machines for efficient training. The trained model was validated on a test dataset to ensure it performed well on unseen data, demonstrating its adaptability to various document types within the financial domain.

TrOCR is a state-of-the-art, transformer-based Optical Character Recognition (OCR) model pre-trained on a large corpus of handwritten text images. To perform inference, pass the cropped image, based on the bounding box for each respective entity, into the TrOCR model to extract text from the image document. Using transformer-based models for object detection and OCR has shown superior performance compared to all other models discussed earlier.

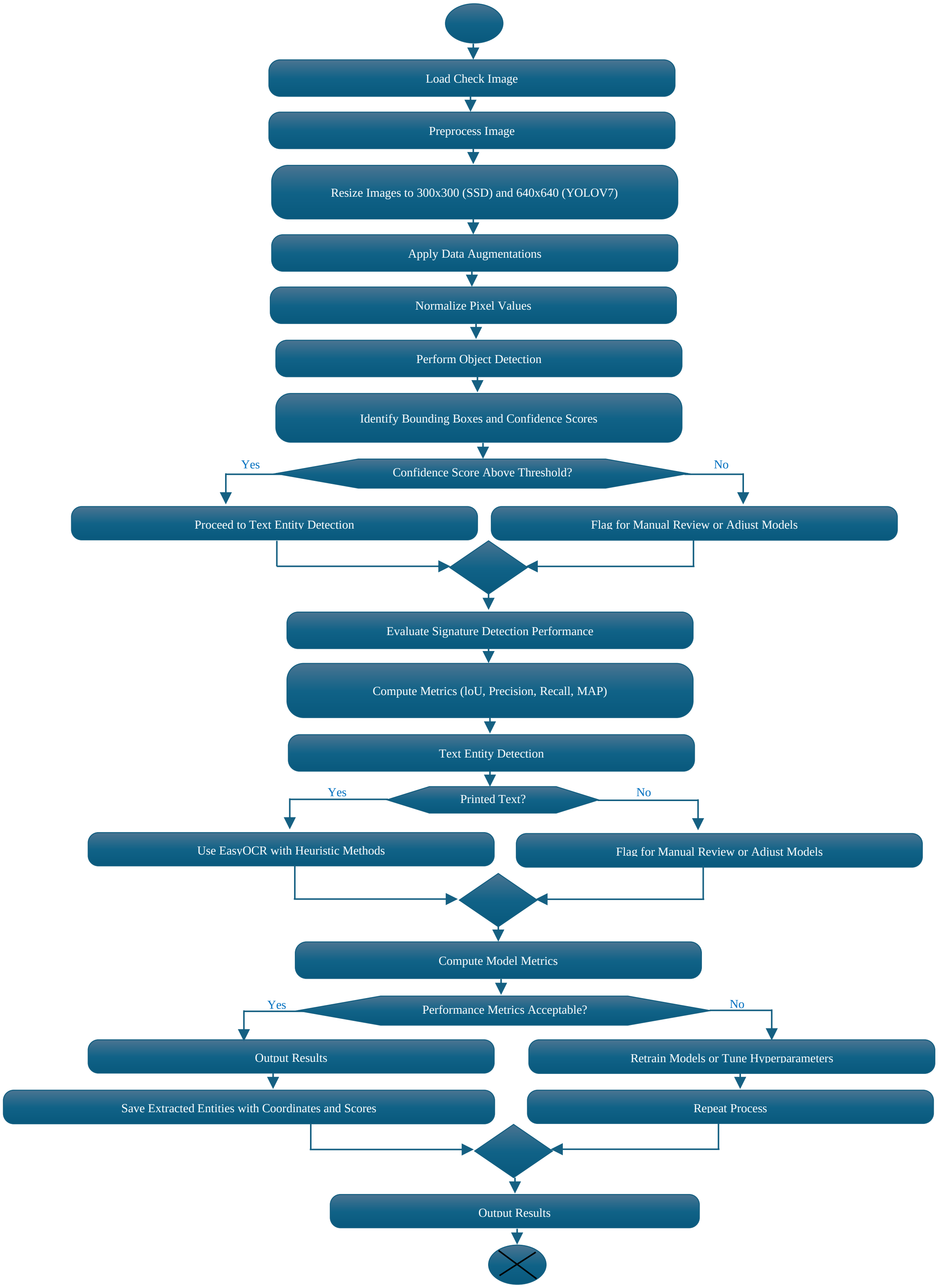
Load Check Image
Preprocess Image
Resize Images to 300x300 (SSD) and 640x640 (YOLOV7)
Apply Data Augmentations
Normalize Pixel Values
Perform Object Detection
Identify Bounding Boxes and Confidence Scores
Yes
Confidence Score Above Threshold?
No
Proceed to Text Entity Detection
Flag for Manual Review or Adjust Models
Evaluate Signature Detection Performance
Compute Metrics (IoU, Precision, Recall, MAP)
Text Entity Detection
Yes
Printed Text?
No
Use EasyOCR with Heuristic Methods
Flag for Manual Review or Adjust Models
Compute Model Metrics
Yes
Performance Metrics Acceptable?
No
Output Results
Retrain Models or Tune Hyperparameters
Save Extracted Entities with Coordinates and Scores
Repeat Process
Output Results

Fig. 2 Process flow diagram of Entity extraction using CNN

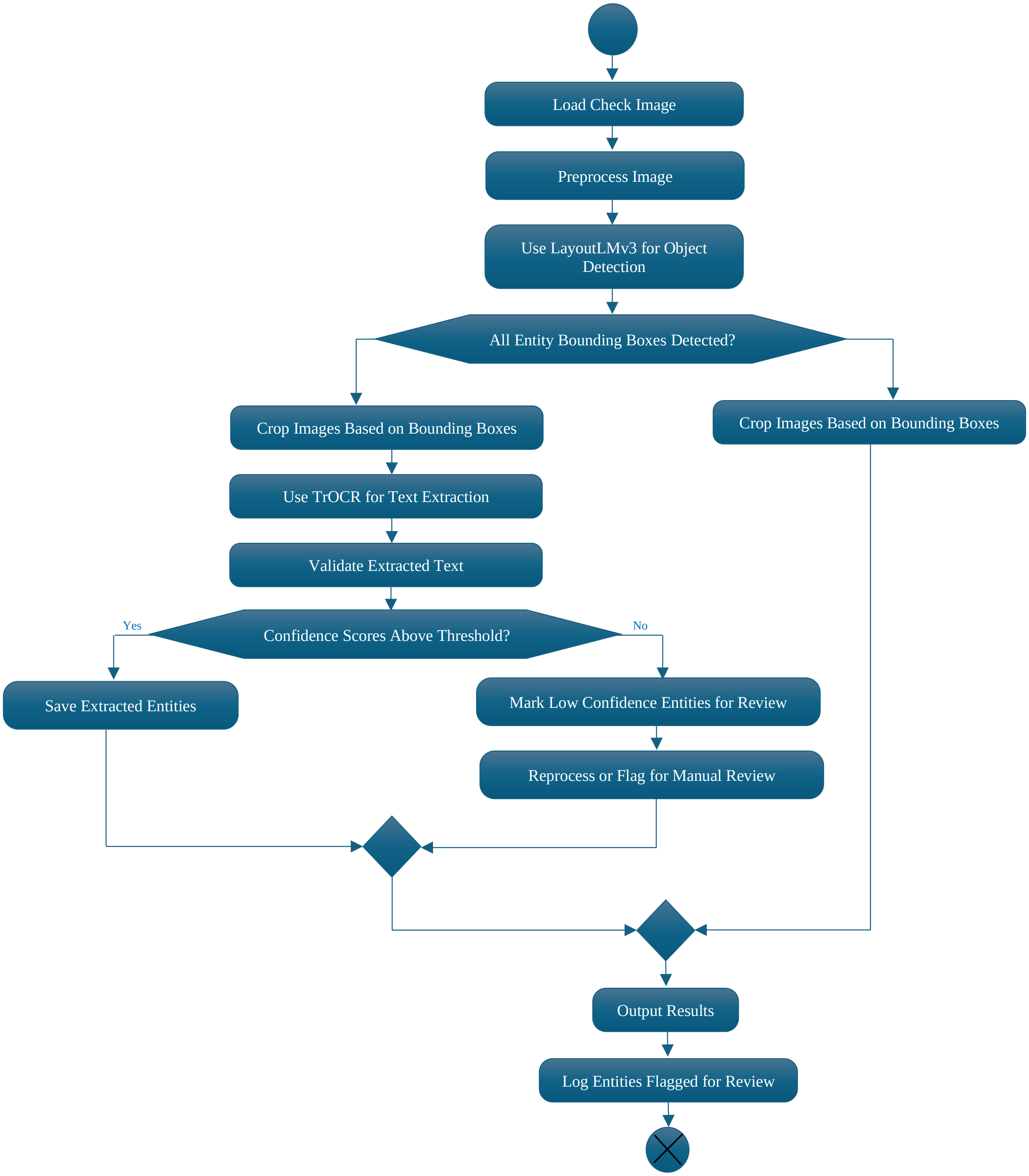
Load Check Image
Preprocess Image
Use LayoutLMv3 for Object Detection
All Entity Bounding Boxes Detected?
Crop Images Based on Bounding Boxes
Crop Images Based on Bounding Boxes
Use TrOCR for Text Extraction
Validate Extracted Text
Yes
Confidence Scores Above Threshold?
No
Save Extracted Entities
Mark Low Confidence Entities for Review
Reprocess or Flag for Manual Review
Output Results
Log Entities Flagged for Review

Fig. 3 Process flow diagram of transformer based model

4. Data Preparation

The documents used in this study were all in PNG format and presented in black and white. To facilitate entity extraction, the data was prepared by annotating relevant entities within the check images. A total of over 1,000 annotated documents were utilized, representing a diverse range of variations, including handwritten checks, printed checks, checks with background text, and documents of varying quality, including some with poor image resolution. To ensure the robustness of the model, approximately 300 check images were set aside as a blind test set. These images were never exposed during the training phase. They were used to test the model's performance on unseen data, ensuring reliability in a real-world production environment.

Text extraction from the images was performed using EasyOCR, a tool built on deep learning frameworks, to perform Optical Character Recognition (OCR). The LabelMe annotation tool was employed for labelling and annotation, and the annotated data was formatted using the CoNLL-2003 standard with an IO schema.

The annotated documents were then randomly divided into training and testing sets, with 70% allocated for training and 30% reserved for testing. This random distribution ensured a fair and consistent assessment of the model's generalization capabilities.

5. Results and Observations

This section discusses the results for different components.

5.1. Heuristic Method Using OCR

Check images though they follow a certain template, they have significant variations in how they appear. The business use cases require extracting seven distinct entities from the check – Date, Amount, amount in text, payer name and address, payee name and address, signature, and MICR line. As discussed, the heuristic model uses cues of the check layout and geometrically traverses through the image to identify and extract the entities. It was observed that the performance of the methodology varied between printed and handwritten checks. This methodology is well suited when all the check contents are printed and legible.

5.2. Signature Detection and Text Entity Detection

A conveyor model for Signature detection and text entity extraction was developed. CNN object detection models, like YOLOv7 and SSD (Single shot detection), were adopted to extract signatures and other text elements. Ensembled results from these two models were used. Data elements other than the signature were passed through the TrOCR transformer for text extraction from the Image. It was noticed that this model performed well for signature detection compared to the other techniques discussed in this paper.

5.3. Transformer-Based Entity Extraction

The transformer model provided significantly superior results with a higher confidence score than the other two models or methods discussed above. The transformer model was specifically fine-tuned on the sample dataset to identify and label entities in the check. The TrOCR transformer model specifically trained on handwritten text produced significantly better performance in extracting handwritten text. The solutions discussed in the paper achieve an overall accuracy of over 94% for text-based entities on check images. These results can be extended to other document types, such as forms, tax, financial, etc.

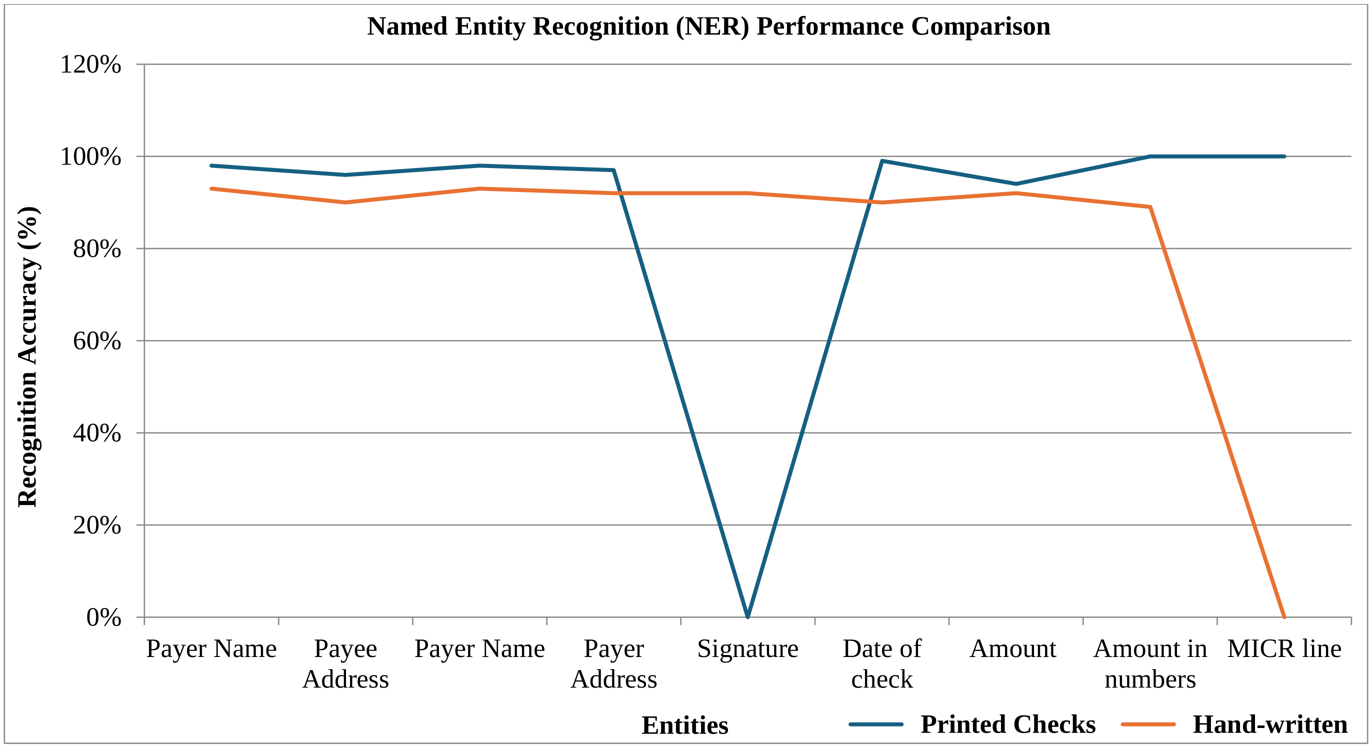

Fig. 4 Printed vs Hand-written model accuracy for entities

Table 1. NER Results

Entities		
Named Entity Recognition (NER)	**Printed Checks**	**Handwritten**
Payer Name	98%	93%
Payee Address	96%	90%
Payer Name	98%	93%
Payer Address	97%	92%
Signature	NA	92%
Date of check	99%	90%
Amount	94%	92%
Amount in numbers	100%	89%
MICR line	100%	NA

6. Discussion

Document extraction is crucial in today's world, where data is golden. This is one area where researchers are still trying to find a solution that would help extract entities from any document type. However, attaining that state is not easy, as the model not only should understand the layout or the landscape of the document but also should intelligently interpret the semantic and relative relevancy of the data points within the context of the document. No single solution would work for a particular document type that exists today. This research paper addresses the gap by adopting different frameworks for different document types and entities with in-check images (handwritten, printed, etc.) to achieve higher recall and accuracy. Check images are processed through all three frameworks in parallel, and the framework with the highest recall and accuracy will be the winner for the respective entity. The motivation behind the concept of the winner framework is that not all models work well on all entities. This was a creative problem-solving approach that would yield highly accurate results on all entities that were extracted. The research presented in this paper is independent and not associated with any organizational collaboration.

7. Conclusion

This research paper explored the adoption of cutting-edge models and their fine-tuning to optimize accuracy and drive impactful business outcomes. Progressing from simple heuristic methods to advanced transformer-based models helped demonstrate how state-of-the-art techniques can be leveraged to solve long-standing challenges. This also highlights that a single model cannot be universally applied to all problems and must be integrated with other models and tailored to meet specific business requirements.

The framework and solutions presented here extend beyond the finance industry, offering a versatile approach that can be applied across various sectors. This research provides valuable insights for future studies on document extraction processes and highlights the importance of addressing its complexities to achieve more efficient and scalable solutions. These modeling patterns contribute towards higher accuracy and reliability by addressing specific challenges associated with different methods. Future work will explore integrating these methodologies with other emerging AI/Generative AI technologies and frameworks to enhance the effectiveness of document understanding further, ensuring continued advancements in building resilient document extraction applications.

References

[1] Ashish Vaswani et al., "Attention Is All You Need," *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, Long Beach, CA, USA, pp. 6000-6010, 2017. [Google Scholar] [Publisher Link]

[2] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao, "YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Vancouver, BC, Canada, pp. 7464-7475, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[3] Wei Liu et al., "SSD: Single Shot MultiBox Detector," *14th European Conference on Computer Vision – ECCV 2016*, Amsterdam, The Netherlands, pp. 21-37, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[4] Guillaume Lample et al., "Neural Architectures for Named Entity Recognition," *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, San Diego, California, pp. 260-270, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[5] Kaiming He et al., "Deep Residual Learning for Image Recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, pp. 770-778, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[6] Pranav Rajpurkar et al., "SQuAD: 100,000+ Questions for Machine Comprehension of Text," *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Austin, Texas, pp. 2383-2392, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[7] Shaoqing Ren et al., "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligences*, vol. 39, no. 6, pp. 1197-1149, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[8] Joseph Redmon et al., "You Only Look Once: Unified, Real-Time Object Detection," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, pp. 779-788, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[9] R. Smith, "An Overview of the Tesseract OCR Engine," *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, Curitiba, Brazil, pp. 629-633, 2007. [CrossRef] [Google Scholar] [Publisher Link]

[10] Alan Akbik et al., “FLAIR: An Easy-to-Use Framework for State-of-the-Art Natural Language Processing,” *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, Minneapolis, Minnesota, pp. 54-59, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[11] John Lafferty, Andrew McCallum, and Fernando Pereira, “Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data,” *ICML '01: Proceedings of the Eighteenth International Conference on Machine Learning*, pp. 282-289, 2001. [Google Scholar] [Publisher Link]

[12] Mohamed Yassine Landolsi, Lobna Hlaoua, and Lotfi Ben Romdhane, “Information Extraction from Electronic Medical Documents: State of the Art and Future Research Directions,” *Knowledge and Information Systems*, vol. 65, pp. 463-516, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[13] David Doermann, “The Indexing and Retrieval of Document Images: A Survey,” *Computer Vision and Image Understanding*, vol. 70, no. 3, pp. 287-298, 1998. [CrossRef] [Google Scholar] [Publisher Link]