Image Text Extraction and Natural Language Processing of Unstructured Data from Medical Reports

Ivan Malashin [1,*], **Igor Masich** [1], **Vadim Tynchenko** [1,*], **Andrei Gantimurov** [1], **Vladimir Nelyub** [1,2] and **Aleksei Borodulin** [1]

[1] Artificial Intelligence Technology Scientific and Education Center, Bauman Moscow State Technical University, 105005 Moscow, Russia; imasich@emtc.ru (I.M.)

[2] Scientific Department, Far Eastern Federal University, 690922 Vladivostok, Russia

[*] Correspondence: i.malashin@emtc.ru (I.M.); vadimond@mail.ru (V.T.); Tel.: +7-926-875-7128 (I.M.)

Abstract: This study presents an integrated approach for automatically extracting and structuring information from medical reports, captured as scanned documents or photographs, through a combination of image recognition and natural language processing (NLP) techniques like named entity recognition (NER). The primary aim was to develop an adaptive model for efficient text extraction from medical report images. This involved utilizing a genetic algorithm (GA) to fine-tune optical character recognition (OCR) hyperparameters, ensuring maximal text extraction length, followed by NER processing to categorize the extracted information into required entities, adjusting parameters if entities were not correctly extracted based on manual annotations. Despite the diverse formats of medical report images in the dataset, all in Russian, this serves as a conceptual example of information extraction (IE) that can be easily extended to other languages.

Keywords: image recognition; natural language processing; named entity recognition; information extraction

1. Introduction

In the era of digitization, the volume of medical documentation continues to grow, presenting challenges in efficient information retrieval and analysis. Medical reports, often available as scanned documents or photographs, contain valuable clinical data essential for healthcare decision making. However, manual extraction and structuring of information from these reports is time-consuming and prone to errors.

Recognition of text [1–3] from images holds significant importance in automating annotation and indexing of visual data, as emphasized in scientific literature exploring methods and algorithms for extracting information from various types of images and videos.

Text information embedded in images and videos is crucial for automatic annotation and indexing, but its extraction presents challenges due to variations in size, style, and orientation, as well as complex backgrounds. Jung et al. [4] addressed the lack of comprehensive surveys by reviewing algorithms for text information extraction (IE), discussing evaluation metrics, and suggesting future research directions.

Text extraction from natural scene images presents a pivotal challenge poised to revolutionize everyday applications such as augmented reality. Comprising detection, localization, enhancement, and recognition stages, this process faces significant hurdles due to variations in text attributes and environmental factors. Zhang et al. [5] categorized and evaluated the latest algorithms, primarily focusing on detection and localization stages, while providing a link to a public image database for evaluation.

Ye et al. [6] provided a comprehensive analysis of text detection and recognition in color imagery, categorizing existing techniques and addressing challenges such as localization, verification, and segmentation. It also discusses specialized issues like degraded text enhancement and multilingual text processing. By comparing benchmark datasets and

evaluating representative approaches, this review aims to identify remaining challenges in the field.

Yin et al. [7] presented a thorough survey of text detection, tracking, and recognition in video, addressing recent advancements and challenges. Unlike previous surveys, it offers a unified framework for understanding detection, tracking, and recognition processes within video text extraction. Special emphasis is placed on text tracking, tracking-based detection, and recognition techniques. The paper also discusses applications, challenges, and future directions in video text extraction, especially from scene and web videos.

A classification-based algorithm for text detection using sparse representation with discriminative dictionaries was proposed by Zhao et al. [8]. This method effectively detects texts of various sizes, fonts, and colors from images and videos, as demonstrated through extensive experiments on different datasets.

A method for text localization and recognition in real-world images was described by Neumann et al. [9]. Departing from traditional feed-forward pipelines, it employs a hypothesis-verification framework and synthetic fonts for training, enhancing robustness to geometric and illumination conditions.

An algorithm for text detection in natural scene images was suggested by Yi [10] addressing the challenges of character appearance and structure variability. It combines a structure correlation approach with structons and correlations to generate discriminative text descriptors. Additionally, it utilizes color decomposition, character contour refinement, and string line alignment for text region localization. Experimental evaluations demonstrate state-of-the-art performance in scene text detection, and character identification.

Ref. [11] introduced a methodology for text extraction from images, encompassing document, scene, and caption text. Leveraging discrete wavelet transform (DWT), the approach accommodates both color and grayscale images. Preprocessing steps are applied to color images, followed by edge detection using the Sobel operator. Morphological operations and thresholding refine the extracted text edges.

The exploration of the application of deep learning [12,13] (DL) in IE [14] (IE) was highlighted by Yang et al. [15], particularly focusing on entity relationship extraction, event extraction, and multi-modal IE. DL techniques have shown superior performance over traditional methods by enabling deeper feature extraction and higher model accuracy. The paper reviews DL's progress in IE tasks, discusses challenges, and suggests future directions like multi-modal extraction and knowledge-enhanced extraction to tackle information overload in the digital age.

Recent advancements in computer vision, particularly with deep learning models, have revolutionized Optical Character Recognition (OCR), but Key IE from documents, a vital downstream task, remains challenging due to underexploitation of semantic visual features. Yu et al. [16] introduced a robust framework combining graph learning and convolution operations to efficiently handle complex document layouts for KIE, yielding richer semantic representations. Extensive experiments demonstrate superior performance over baseline methods, addressing the critical need for efficient and accurate KIE.

Image-based Text Extraction is vital for extracting valuable information from images. Optical Recognition Systems like Tesseract employ algorithms for text detection, localization, and recognition. Revathi et al. [17] proposed using image processing techniques to enhance text extraction, offering adjustable parameters for optimal results. This method promises significant importance in automation, ensuring efficient extraction of text from various backgrounds.

The integration of text recognition from images with subsequent NLP processing [18] has been widely explored in various sources, serving as a crucial step towards extracting meaningful insights from visual data. Wang et al. [19] proposed a hybrid network method for English word segmentation, while Ref. [20] explored the applications of NLP in materials science. In the medical field, for instance, Ref. [21] explored the intersection of computer vision and NLP to aid vision-impaired individuals.

Ref. [22] demonstrated the effectiveness of a hybrid approach combining OCR with natural language processing (NLP) to extract clinical information from scanned colonoscopy and pathology reports. Conducted at the Cleveland Clinic and the University of Minnesota, the retrospective study compared manual data extraction with the OCR/NLP hybrid approach across various variables. Results showed high accuracy rates, exceeding 95% for detecting various clinical parameters, including polyps, adenomas, and ECT. This innovative hybrid technology offers a reliable method for obtaining critical clinical data from scanned reports, facilitating streamlined quality management and risk assessment in colorectal cancer screening.

As electronic health records (EHRs) become more prevalent, extracting valuable information from unstructured text data poses a significant challenge. IE using NLP techniques can automate this process, yet its utilization in clinical research remains limited. Ref. [23] aimed to understand this under-utilization and propose strategies to enhance the integration of NLP in EHR-based clinical studies.

Viani et al. [24] addressed the need to extract structured clinical information from Italian medical reports. They proposed an ontology-driven NLP approach to convert unstructured text into queryable data while preserving semantic relations and focus is on answering two key questions: Can automated extraction produce structured data? Furthermore, can semantic relations be preserved? They utilized a domain ontology to guide the extraction process, aiming for extensibility and language-independence, with Italian medical reports.

Over the last decade, Deep Learning (DL) has revolutionized NLP, particularly in medicine, by effectively handling unstructured textual data. DL eliminates the need for manual feature selection, significantly improving performance in tasks like IE from medical texts. However, applying DL to medical language poses challenges due to specialized vocabulary and syntax complexities, requiring adaptations of general language embeddings. Ref. [25] underscored DL's transformative impact on NLP methodologies in medicine, complementing previous research in medical informatics.

Our research aims to fill the gap in knowledge by developing an integrated approach for automated information extraction from images of medical documents. This gap exists due to the lack of comprehensive studies that combine adjusting OCR parameters to extract unstructured personal/financial data contained in medical reports and structure them using NLP. By addressing this gap, our study seeks to contribute to the advancement of automated data extraction methodologies in healthcare, enabling more efficient processing of medical information.

To address our primary research challenge, we aimed to develop an adaptive model capable of efficiently extracting text from medical report images. We achieved this by employing a genetic algorithm to fine-tune OCR hyperparameters, maximizing the length of the extracted texts. Next, we categorized the extracted information into required entities using NLP techniques. If an entity was not correctly extracted according to the manually annotated dataset for this study, we adjusted the parameters accordingly. This adjustment involved passing the image data (see Appendix A) and the optimal OCR recognition parameters through a neural network. Here, the image data served as inputs, while the optimal OCR recognition parameters served as outputs. The resulting neural network was then applied to new images to maximize the extraction of the required entity.

The structure of our article is as follows: Section 2.1 describes the approach of manual information extraction for the required entities, Section 2.2 presents the automated process, and Section 2.3 discusses NLP-based entity extraction. In Section 3.1, we provide an analysis of manual statistics used as dataset annotations, while Section 3.2 compares the correlation of automatically extracted image metadata, revealing that the optimal parameter for maximizing information extraction is maximizing the number of recognized characters through hyperparameter tuning using GA. Section 3.3 describes this model, while Section 3.4 presents comparative results before and after its application for the

considered entities. Finally, Section 4 outlines the limitations of the approach and prospects for future work.

2. Materials and Methods

2.1. Manual Quality Check of Reports

For the research, an initial dataset of folders containing .pdf scans and photographs of medical documents was utilized. To standardize the format and ensure consistency of the input data, all images were converted to .jpg format. The diagram illustrating the research methodology is depicted in Figure 1.
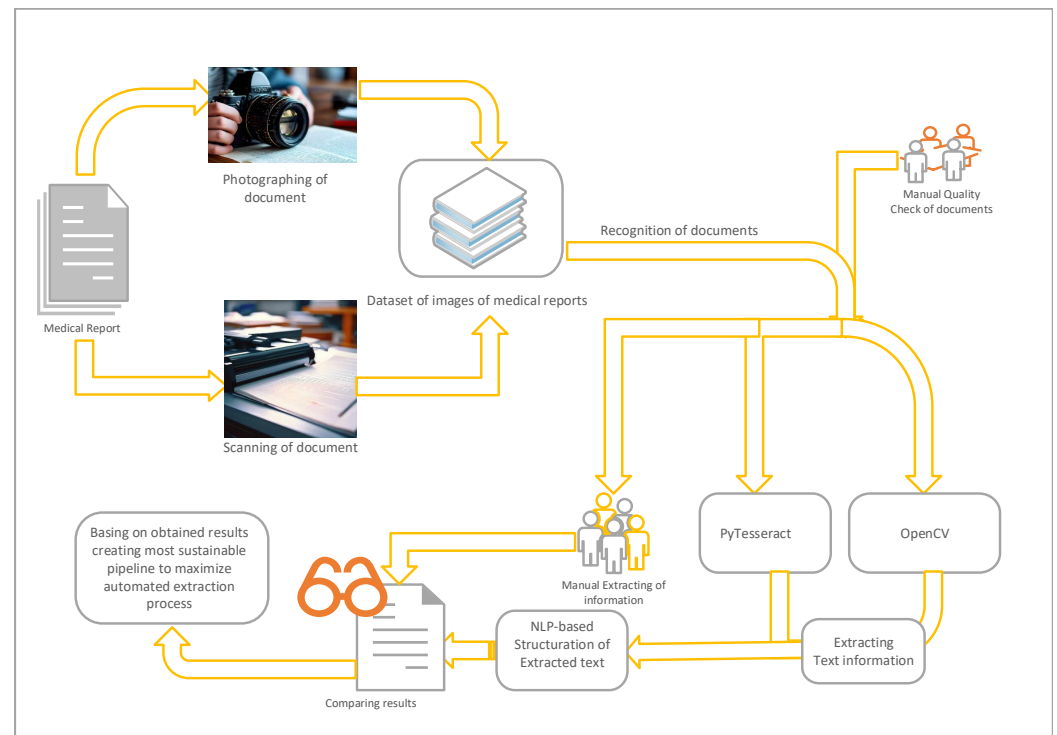

Figure 1. Diagram illustrating the research methodology employed in the study.

Manual evaluation of image darkness was conducted manually, categorizing images as either dark (1) or not dark (0), to identify and rectify insufficient or excessive lighting conditions that could impact text recognition accuracy.

Additionally, an assessment of image blurriness was conducted, categorizing images as blurred (1) or not blurred (0), to understand the impact of blurriness on text extraction. This criterion was instrumental in refining preprocessing techniques aimed at improving text clarity before being processed by the pytesseract and easyOCR libraries

The presence of handwritten text was manually evaluated, labeled as 1 if present and 0 if absent, considering potential challenges it may pose for text recognition libraries. This aspect of the methodology aided in identifying scenarios where additional processing techniques, such as segmentation, could enhance recognition accuracy.

Entities from the dataset of medical report images were manually extracted by two annotators independently across the entire dataset. Subsequently, their work was evaluated through a character-level comparison of the entities identified by each annotator. The outcome of this comparison revealed a match rate of 1661 out of 1674 printed medical reports for all required entities, including the cost of medical services, payment dates, taxpayer identification numbers, and medical institution license numbers. For cases where discrepancies occurred, additional verification was conducted to address inaccuracies.

2.2. Automated Processing of Reports

Automated processing involved recognition through PyTesseract and easyocr. OCR encounters challenges with Russian language support, often substituting English characters for unrecognized Russian ones.

Our work also involves developing a method for automatically obtaining metadata from images. Image sharpness was obtained using the Laplace operator. The Laplace operator computes the second derivative of image brightness, enabling the detection of areas with rapid intensity changes. After processing, each pixel in the resulting image represents an approximation of the second derivative of brightness in both directions. Subsequently, the variance of the obtained image values is calculated. A high variance value indicates high image sharpness. Figure 2a,c visually illustrate an image used for variance calculation. An unclear image (Figure 2) had a variance of 774, while a sharp one had a variance of 5277.

Image sharpness obtained using the Sobel operator to detect edges in an image by highlighting brightness gradients at each pixel. After processing, each pixel in the resulting image represents an approximation of the brightness gradient in a specified direction. The sum of the square roots of horizontal and vertical gradients is then computed to estimate overall image sharpness. A higher value indicates greater image sharpness. Figure 2b,d visually depict an image used for this estimation. For clarity, the sharpest ($16{,}182 \times 10^6$) and blurriest (125×10^6) images, the latter of which is unreadable to the human eye, were selected.

(**a**)　(**b**)　(**c**)　(**d**)

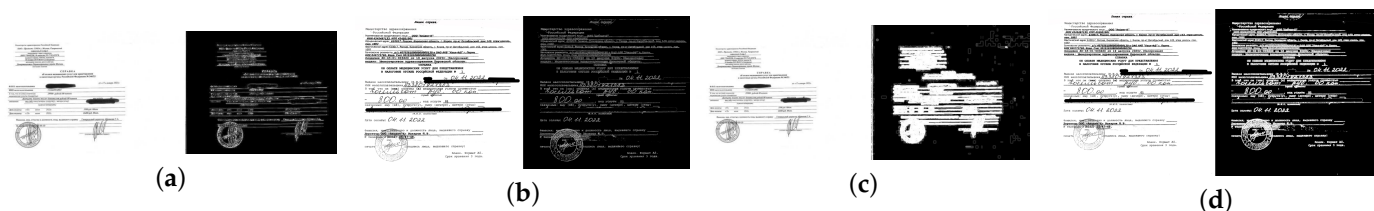**Figure 2.** (**a**,**c**) Processing of a blurry image with the Laplace and Sobel operator, respectively: on the right—the original image, on the left—the image processed using the Sobel operator. (**b**,**d**) Processing of a sharp image with the Laplace and Sobel operator, respectively: on the right—the original image, on the left—the image processed using the Sobel operator.

2.3. NLP-Based Text Structurization

During the document recognition process, we often encountered a jumble of letters and symbols in obtained texts. As a result, it was imperative to structure this information coherently, facilitating its comprehension and further analysis. NLP [26,27] was employed to extract specific information from the recognized text. This involved structuring the extracted data to identify key elements such as the organization's Tax Identification Number (TIN), taxpayer identification number, license number, cost of medical services (sometimes written in words or digits), and dates of medical service payments (also variably represented in words or digits), which poses several challenges.

Firstly, the variability in formatting and presentation across different documents can lead to inconsistencies in the structure and content of the extracted information. For instance, the TIN, taxpayer identification number, and license number may be presented in different formats or locations within the document, making it challenging to consistently identify and extract this information accurately.

Secondly, the presence of handwritten or poorly scanned text can introduce errors or inaccuracies during the OCR (Optical Character Recognition) process, particularly when dealing with numerical values or dates that may be misinterpreted due to irregular handwriting or smudging.

Furthermore, the variation in language and writing styles can complicate the extraction process, especially when dealing with numeric values written in words or dates expressed

in text format. NLP models may struggle to accurately interpret such information, leading to potential errors in extraction.

Additionally, the absence of standardized templates or document structures [28] adds another layer of complexity, as the relevant information may be dispersed throughout the document or presented in unconventional formats, requiring sophisticated NLP algorithms to identify and extract the desired data accurately.

To extract the required entities, Named Entity Recognition (NER) algorithms from the PullEnti [29,30] and Natasha [31,32] Python libraries were employed. These algorithms were used to identify patterns associated with the specified entities effectively [33].

Monetary values, comprising currency symbols, numerical amounts, and associated terms like "price", "cost", or "payment", are common patterns found in medical reports. To extract such entities, we leveraged PullEnti, utilizing entity extraction with specified labels to target and capture relevant information.

Similarly, for extracting the Taxpayer Identification Number (TIN) and organization identification numbers, as well as the license numbers of medical institutions, PullEnti was harnessed to identify patterns specific to these entities within the document text. Patterns related to TINs, typically composed of 12 digits encompass various numerical sequences with specific lengths and formats, alongside associated labels like "Tax ID", "TIN", or "Identification Number". License numbers for medical institutions usually consist of alphanumeric sequences interspersed with dashes, often accompanied by adjacent terms such as 'license'.

For extracting payment dates, the DatesExtractor method from the Natasha library was applied to recognize temporal expressions [34,35] and date formats within the text data. This involves identifying patterns associated with dates, such as month-day-year formats, numerical sequences, and textual representations of dates.

3. Results

3.1. Manual Labeling Statistics

The dataset consisted of 2041 documents, primarily in jpg/jpeg format, with 367 containing handwritten text. Among these, 1674 documents exclusively featured printed text. Despite variations in document quality, including scans, low-quality photographs, tilted images, and obscured text, each document was expected to contain essential information such as TINs, organization details, service costs, payment dates, and medical institution names and addresses. Our analysis focused on accurately identifying these entities within the printed text documents.

The final outcome of the manual image classification process is depicted in Figure 3a, showcasing all conceivable combinations of document pairs. This visual representation stage enables a comprehensive assessment of the manual classification outcomes for all images. Following this, our analysis exclusively scrutinized documents featuring printed text, as delineated in Figure 3b.
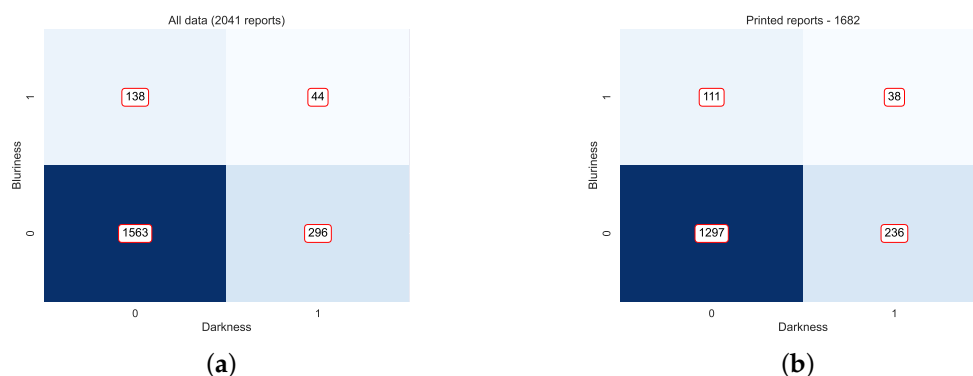
Figure 3. (**a**) The number of combinations of readable/dark reports among all available; (**b**) only among printed reports.

3.2. Correlation Analysis of Parameters

Figure 4 illustrates correlation matrix [36] for automatically identified parameters in medical reports recognition it provides quantitative metrics for assessing document quality, aiding quality control and assurance processes. The visualization of these parameters offers a comprehensive overview of document features, enabling rapid anomaly detection.
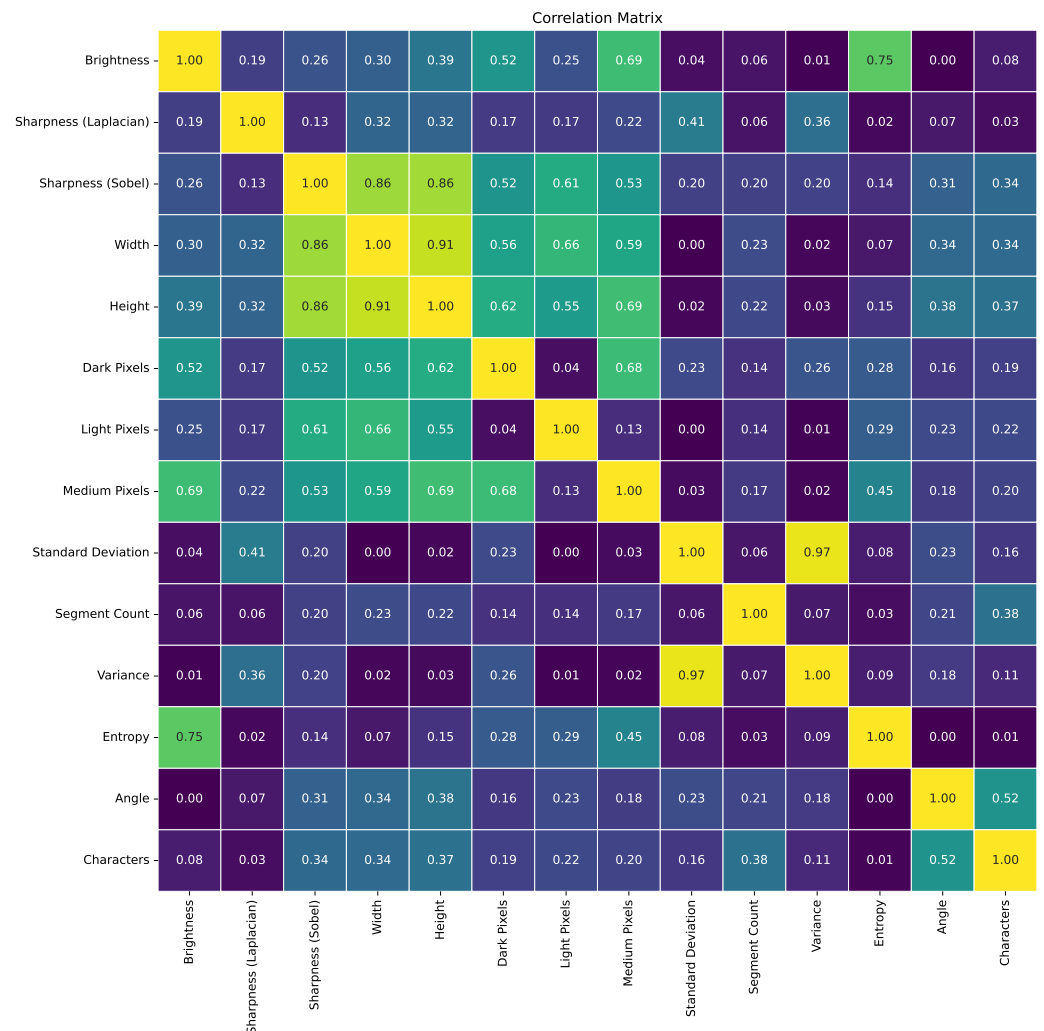

Figure 4. Correlation matrix of automatically detected parameters in medical reports recognition.

Particularly noteworthy is the strong correlation between entropy and brightness (correlation coefficient = 0.75). This correlation suggests that variations in document brightness levels influence the degree of chaos or diversity in brightness gradients [37] within the document, thereby affecting entropy. Possible reasons for this correlation include variations in document background complexity, lighting conditions during scanning, and text font styles, all of which can influence both brightness levels and entropy.

Furthermore, the angle of inclination of the document (presumably representing skewness) demonstrates a considerable correlation with the number of recognized characters (correlation coefficient = 0.52). This correlation indicates that the skewness of the document affects the accuracy of character recognition, as skewed documents may pose challenges for OCR algorithms.

Regarding sharpness measured via the Sobel operator, it exhibits moderate correlations (correlation coefficients > 0.2) with each of the other parameters. This observation suggests that sharpness, as captured by the Sobel operator, is influenced by various document characteristics, such as brightness, size, pixel distribution, and segmentation count.

Conversely, the parameter with the lowest correlation with other parameters is segment count. This finding indicates that the segmentation count, which represents the number of distinct segments within the document, is relatively independent of other document features captured by the analyzed parameters.

Therefore, the analysis leads to the conclusion that the most influential factor after OCR recognition is the quantity of recognized characters. As a result, we set out to develop an adaptive framework aimed at adjusting OCR parameters to maximize IE.

3.3. Adaptive Model for Document Recognition

Automatic entity extraction using NLP shows promise for document recognition. However, disparities between automatically and manually extracted entities highlight the necessity for an adaptive model. Such a model would adjust document parameters to enhance entity recognition accuracy based on recognized characteristics.

The meticulous adjustment of OCR parameters is essential for accurate IE. We propose an approach based on using a genetic algorithm (GA) to optimize parameters for OCR. Our goal is to maximize the number of recognized characters in document images, thereby increasing the amount of information available for entity extraction using NLP methods.

We define a set of OCR parameters, including beam width, batch size, contrast threshold, and others, and utilize a genetic algorithm to search for optimal values for these parameters. To evaluate the quality of each parameter set, we load the document image, convert it to grayscale, apply a threshold value to improve text recognition quality, and use the PyTesseract and easyOCR library to recognize text on the image. The evaluation metric is the number of recognized characters in the image.

Through an iterative process of the GA, we select the best parameter sets based on their quality evaluation. These best parameters can then be used to enhance the efficiency of OCR and subsequent entity extraction from documents using NLP. Figure 5 depicts the schematic representation of the proposed approach where "int" indicates the selection of integer parameters, while "float" represents floating-point parameters.
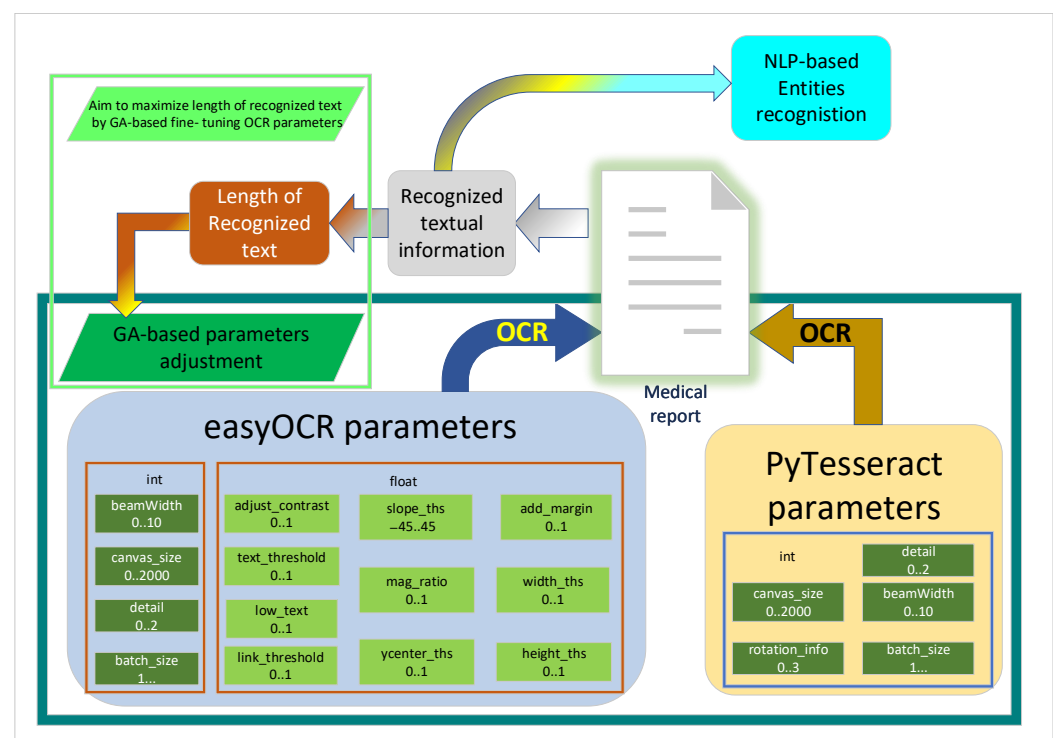

Figure 5. Illustration of the proposed approach utilizing GA for optimizing OCR parameters to maximize character recognition in document images, facilitating enhanced entity extraction through NLP techniques.

After applying GA to fine-tune the optimal hyperparameters for OCR medical document recognition, aiming to maximize the extraction of recognized characters, we constructed a dataset. This dataset contains information about the optimal parameters selected during OCR recognition and the original characteristics of the images. Detailed descriptions of these characteristics are provided in the Appendix A. Subsequently, a neural network (NN) was employed, taking the image's input parameters and outputting the optimal OCR recognition parameters identified in the GA-based step. The architecture of the proposed neural network is depicted in Figure 6.
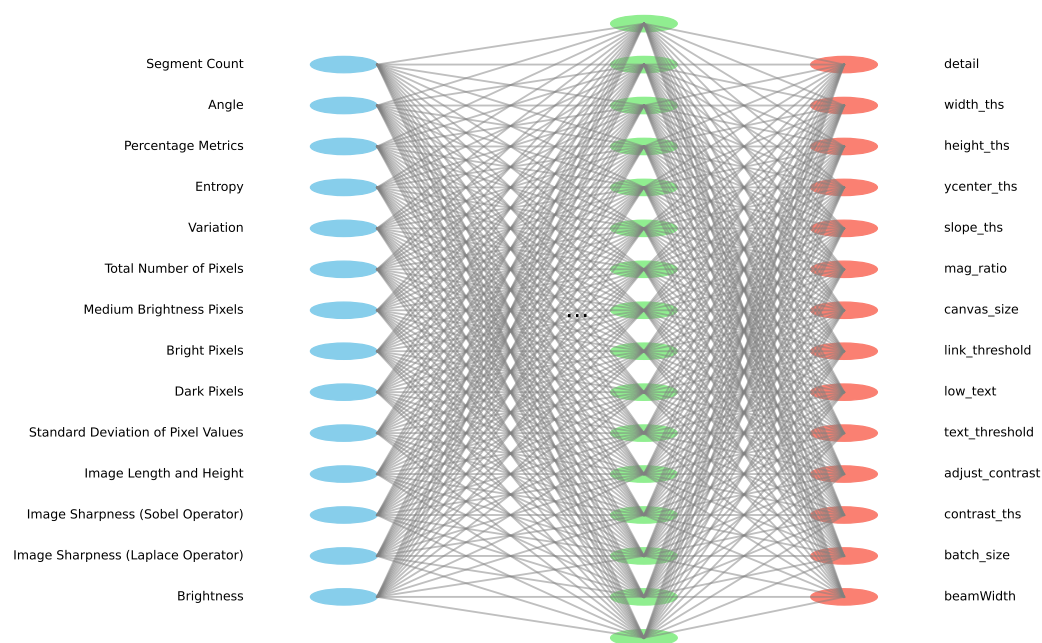

Figure 6. Illustration of the possible structure of the proposed NN.

The objective of this stage is to preselect OCR parameters that can most effectively extract information from medical document images in the future, particularly on new datasets. For each new medical document image where necessary information could not be obtained using conventional methods, all input data described in the appendix is gathered. Then, these data are fed into the pre-trained NN, which provides optimal OCR parameters. Subsequently, the recognition process is conducted once again, utilizing these optimal parameters, in an attempt to extract the required information accurately.

General scheme of an adaptive model for adjusting image parameters and OCR settings to maximize IE is shown on Figure 7. The process entails gathering diverse images of medical report documents. Subsequently, the GA-based fine-tuning technique is applied to OCR hyperparameters to enhance character recognition in images. Leveraging the information gathered on OCR parameters and input image characteristics, a neural network model is employed to optimize OCR settings accordingly. Following this optimization, entity recognition is performed on the extracted texts using NLP techniques. If certain entities remain unrecognized, a feedback mechanism is employed. This mechanism enables the model to adapt image parameters and OCR settings based on learning from new data and feedback from previous recognition attempts. Additionally, regular assessments of the model's performance on new data are conducted to update parameters and refine the adaptive tuning algorithm iteratively. By continuing this iterative process of training and adaptation, the aim is to achieve maximum IE from document images. Ultimately, such a system has the potential to reduce manual effort in entity extraction from document image.
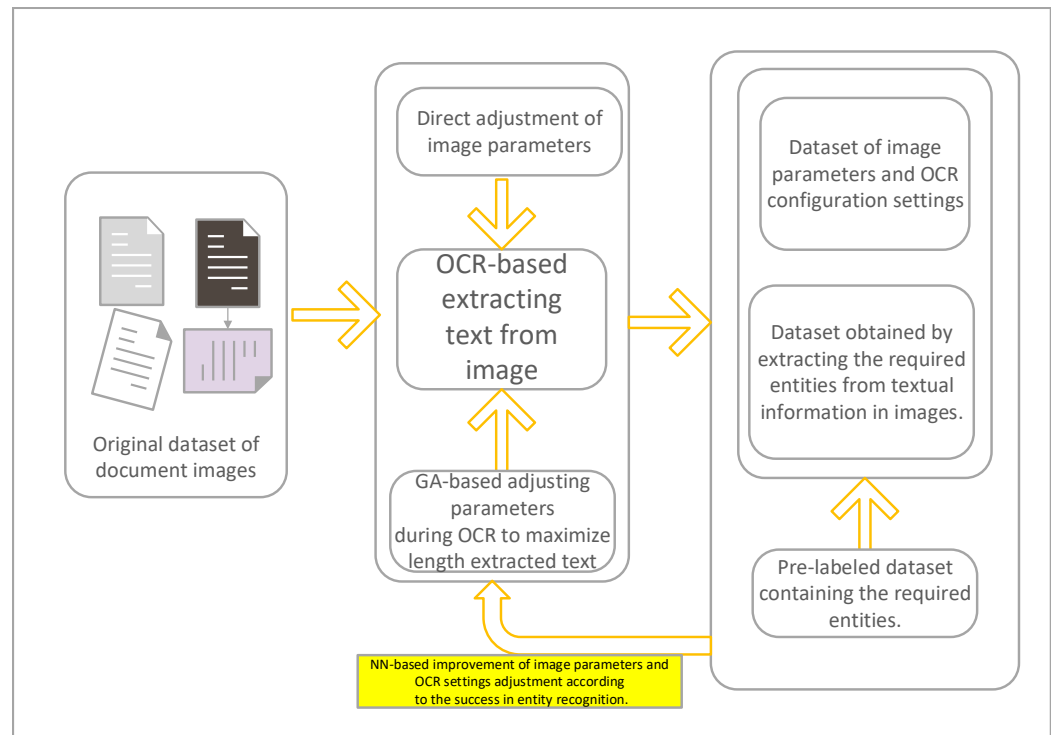
Figure 7. Schema showcasing the adaptive model's functionality in adjusting image parameters and OCR settings to optimize IE from document images.

Based on the analysis of correlations between image parameters and the quality of entity recognition, the model identifies the necessary parameter corrections. For example, if the tilt angle of the document hinders accurate text recognition, the model may decide to automatically straighten the image. Conversely, if the skew angle of a document impedes accurate character recognition, the model can rectify this by aligning the document horizontally [38]. Similarly, if a document appears too dark or overexposed, adjustments can be made to optimize brightness and contrast. Furthermore, if the document is blurry, appropriate transformations can be applied to enhance IE.

3.4. Adaptive Model Outcomes

The bar chart visualization presented in Figure 8 illustrates the comparison between the recognized TINs for organizations and taxpayers before and after the implementation of the adaptive model. This model enhance the accuracy of TIN recognition by adjusting document parameters during OCR.

Additionally, we conducted a comparison of recognized payment dates and payment amounts before and after the application of the adaptive model. This comparison was performed on a dataset consisting of 1382 documents, where both the payment date and amount were mentioned only once within each document. The purpose of this analysis was to assess the effectiveness of the adaptive model in improving the accuracy of payment date and amount recognition. Results are shown in Figure 9a,b.
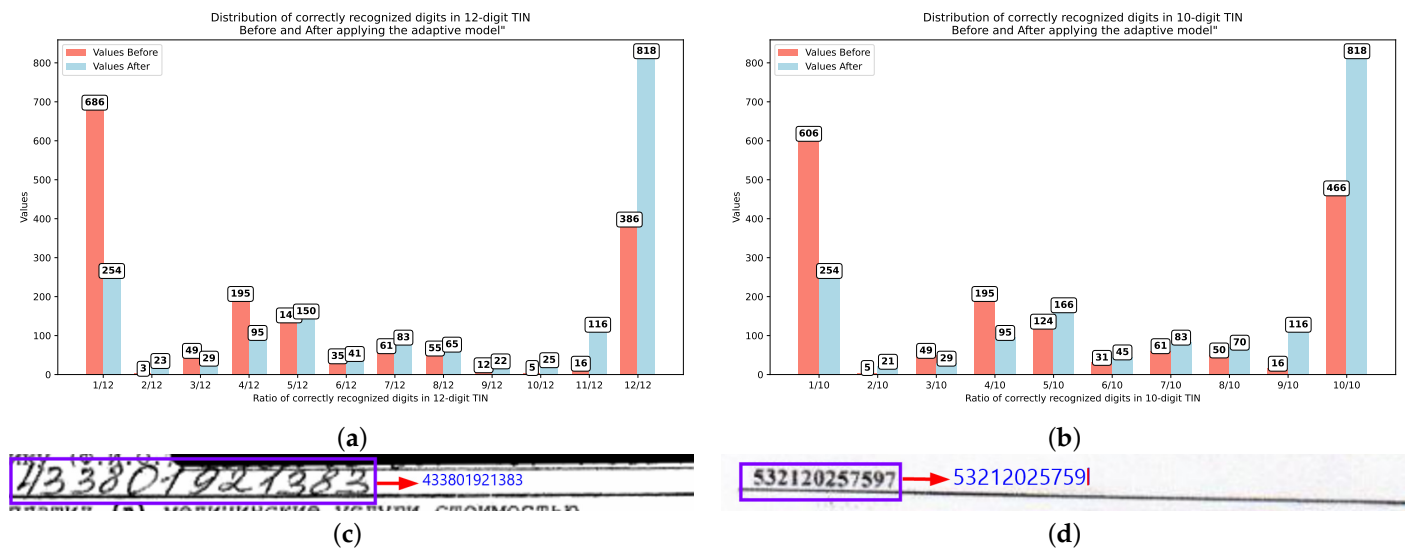
Figure 8. Comparison of recognized TINs before and after the application of the adaptive model for (**a**) 12-digit TINs and (**b**) 10-digit organization TINs. The bar chart visualizes 1682 printed reports, where each pair of adjacent bins represents the number of correctly recognized digits before (red) and after (blue) the application of the adaptive model. (**c**) An example of successfully recognized Taxpayer Identification Number (TIN) where all 12 digits are correctly identified. (**d**) In another instance, only 11 out of 12 digits are correctly recognized. One possible reason for this discrepancy could be due to the tilt of the medical report image.

It is worth noting that this comparison was conducted specifically for cases where the payment date and amount appeared only once within each document. There were 1383 instances out of 1682 printed reports where the payment date and amount appeared only once within each document. This approach was chosen due to the complexity of accurately identifying multiple instances of these elements within a single entity. By focusing on documents with singular occurrences of payment dates and amounts, we aimed to ensure a more precise evaluation of the adaptive model's performance in enhancing recognition accuracy.

Extracting addresses of medical institutions using NER presents several challenges. Firstly, medical reports often contain unstructured text, making it difficult for NER models to accurately identify address entities amidst other information. Additionally, variations in address formats and abbreviations further complicate the extraction process.

Furthermore, the context in which addresses appear within medical reports can vary, affecting their recognition. For example, addresses may be embedded within narrative text, tables, or headers, requiring robust NER models capable of handling diverse document layouts.

Another challenge involves the comparison of extracted addresses with pre-labeled data. Inconsistencies between the extracted addresses and the ground truth annotations may occur due to differences in formatting, spelling variations, or missing information.

Before the implementation of the adaptive model, the distribution of recognized address entities exhibited inconsistencies. However, following the integration of the adaptive model into the OCR process, significant improvements were observed, as depicted in Figure 9c,d.
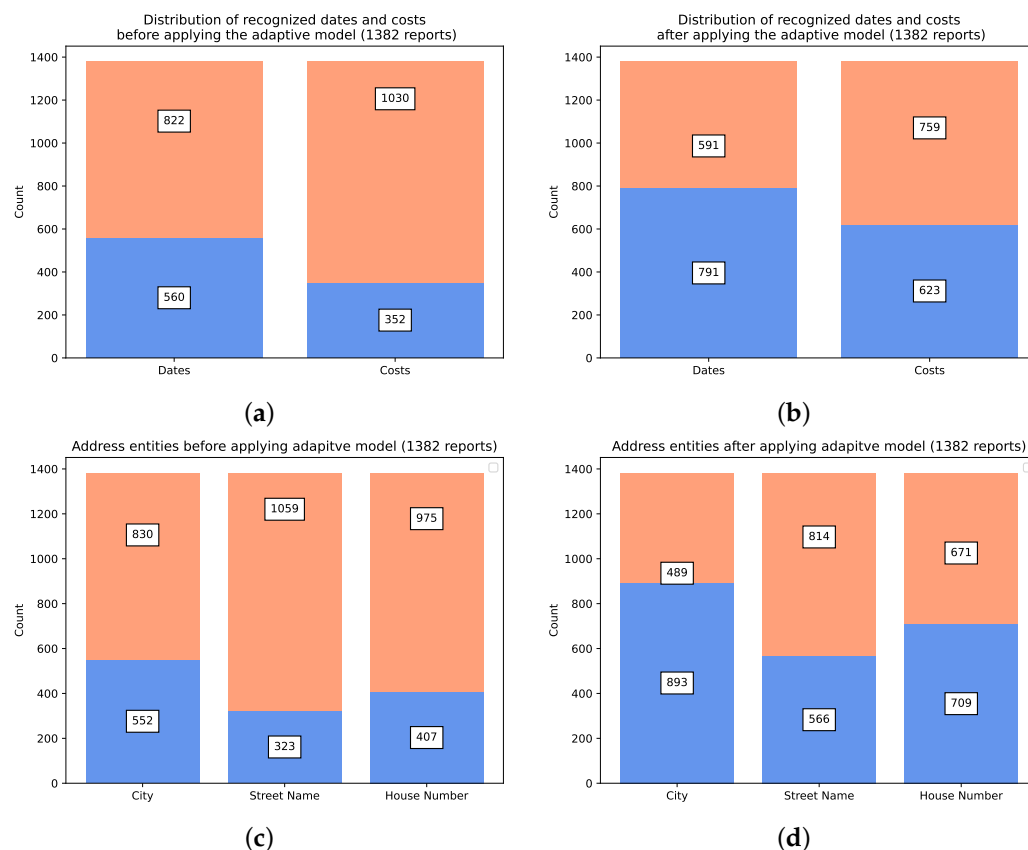
Figure 9. Comparison of recognized payment dates costs (**a**) before and (**b**) after the application of the adaptive model for 1382 documents. Distribution of recognized address entities (**c**) before and (**d**) after the application of the adaptive model to the OCR process. Red corresponds to the number of incorrectly determined entities, while blue indicates correct ones.

4. Discussion

In today's world, artificial intelligence (AI) progress in NLP and computer vision (CV) enable systems to automatically analyze texts, images, and audio data. This progress enhances efficiency and accuracy across various sectors like healthcare [39], finance [40], and education [41], facilitating faster and more precise decision making. Ranaldi et al. [42] questions the direction of AI progress, particularly the shift towards auto-epistemic logic [43] by statistical learners by exploring fundamental issues in AI analysis, including symbolic structure and strategic reference points, and traces the evolution of knowledge representation.

Document recognition using ML algorithms has emerged as an area of research, enabling automated extraction of structural components and semantic information from diverse document layouts, facilitating efficient document analysis and retrieval in large-scale digital libraries. Paass et al. [44] suggested to leveraging ML approaches, such as support vector machines and Conditional Random Fields, offers promising solutions to adaptively handle diverse document layouts and evolving structural features over time. Ghazal et al. [45] presented a handwritten document recognition system using convolutional neural networks (CNNs). It aims to assist visually impaired users and automate data entry tasks.

Central to our investigation was the meticulous deployment of NLP algorithms for entity recognition, coupled with the development of adaptive models tailored for document parameter adjustment. These methodologies were instrumental in enhancing the precision and efficiency of IE processes, thereby accentuating their practical applicability in real-world settings. NER is vital for extracting key information from vast textual data sources.

However, this study is not devoid of its limitations. Challenges encountered during the IE process, such as inaccuracies in entity recognition and the handling of complex

document formats [46], underscore the need for continued refinement and optimization of our models. Addressing these constraints is imperative to bolster the robustness and versatility of our methodologies.

Nevertheless, the implications of our research are profound. The developed models and techniques harbor the potential to revolutionize document processing workflows [47] across diverse domains, including healthcare institutions and financial organizations. By automating labor-intensive tasks and minimizing errors, these advancements pave the way for enhanced efficiency and data quality.

One potential limitation of proposed GA-based approach is the complexity [48,49] associated with optimizing hyperparameters in OCR. While using GA allows for efficient parameter tuning to maximize length of text extraction, this process can be time consuming [50,51] and require significant computational resources, especially when dealing with a large volume of images. Additionally, optimal hyperparameters may vary depending on various factors, such as image quality, document type, and lighting conditions, which can further complicate the optimization process. Another limitation is the need for a large volume of annotated data for model training and performance evaluation, particularly when using ML methods for parameter tuning. Thus, successful implementation [52] of the approach requires consideration of these limitations and the development of effective strategies to overcome them.

Another aspect to consider is the potential application of alternative heuristic algorithms for this approach, such as Particle Swarm Optimization (PSO) [53] and Simulated Annealing [54]. These algorithms offer different strategies for parameter optimization in OCR. For example, PSO optimizes solutions by iteratively adjusting a population of candidate solutions based on their fitness. Simulated Annealing mimics the annealing process in metallurgy, gradually decreasing the temperature to explore the solution space effectively [55]. By incorporating these algorithms, it is possible explore a broader range of optimization techniques and potentially achieve even better results in text extraction from medical report images. However, each algorithm has its strengths and weaknesses, and selecting the most suitable one depends on factors such as problem complexity, computational resources [56], and the specific objectives of the study. Further research could investigate the performance of these alternative algorithms and compare their effectiveness in optimizing OCR parameters for maximizing length of extracted text.

While adapting document processing techniques might seem like an intuitive solution, it may not necessarily address intrinsic errors inherent to the NLP tools. These tools undergo complex processes to interpret and extract meaning from text, including syntactic and semantic analysis, entity recognition, and more. Any inaccuracies stemming from these processes could propagate through the entire workflow, impacting downstream tasks such as document processing. For instance, errors in entity recognition, particularly with respect to formatting inconsistencies, could lead to discrepancies in the information extracted from the documents. These errors might persist even with adaptations in document processing strategies if not directly addressed at the NLP tool level.

Therefore, a thorough examination of the relationship between NLP tool performance and adaptive OCR model performance is essential. This could involve conducting targeted evaluations to isolate the sources of errors [57] and discern whether they originate from the NLP tools themselves or are exacerbated by interactions with the OCR model. Additionally, exploring potential synergies between the NLP tools and OCR model could yield insights into optimizing overall system performance. Strategies such as fine-tuning NLP models based on feedback from OCR outputs or integrating contextual information from OCR-detected regions could enhance the accuracy and robustness of IE pipelines [58].

Looking ahead, future research endeavors could delve into the integration of state-of-the-art neural network architectures for image segmentation and text recognition tasks. By leveraging advanced deep learning techniques, such as convolutional neural networks [59–61] (CNNs) for image segmentation and recurrent neural networks [62–64] (RNNs) for text

recognition, researchers can explore more sophisticated methods for document processing and IE.

5. Conclusions

The study highlighted the approach of IE from images of documents by dynamically adjusting OCR parameters through an adaptive model, followed by the extraction of required entities using NLP techniques. The efficacy of these methods was demonstrated through the comparison of the extracted entities like organization identification numbers, TINs, license numbers, payment amounts, and payment dates. Such application of an adaptive model in OCR image recognition can be beneficial for organizations managing tax deductions for employees, especially when dealing with numerous documents, or for tax authorities handling a large volume of such unstructured documents. This approach enables the reduction of laborious manual extraction of such information, streamlining the process significantly.

Our findings highlight the potential of automated document processing workflows to enhance efficiency, accuracy, and data quality across various sectors, including healthcare and finance. By automating labor-intensive tasks and minimizing errors, these advancements have the capacity to revolutionize document processing practices and streamline operational workflows.

However, this study also underscores the need for continued refinement and optimization of adaptive model to address challenges such as inaccuracies in entity recognition and the handling of complex document formats. Future research endeavors should focus on advancing NLP techniques, integrating ML algorithms for document classification and feature extraction, and exploring novel approaches to enhance IE accuracy.

In summary, our study contributes valuable insights to the field of document recognition and IE, offering both achievements and avenues for improvement. Through ongoing research and development efforts, we can further harness the potential of NLP methodologies to drive innovation and progress in document processing and related domains via ML techniques.

Appendix A. Image Characteristics

This section provides explanations of the key image characteristics utilized in this study:

1. Brightness: Ensures adequate illumination for accurate text recognition.
2. Image Sharpness (Laplace Operator): Detects areas with rapid intensity changes, crucial for identifying text clarity.
3. Image Sharpness (Sobel Operator): Highlights brightness gradients to enhance edge detection and overall text sharpness.
4. Image Length and Height: Provides dimensions for contextualizing image content.

5. Standard Deviation of Pixel Values: Quantifies pixel value variability, indicative of image clarity.
6. Dark Pixels: Identifies low-intensity areas, aiding in contrast enhancement for text extraction.
7. Bright Pixels: Highlights high-intensity areas, assisting in text differentiation from backgrounds.
8. Medium Brightness Pixels: Captures text within a balanced brightness range for optimal recognition.
9. Total Number of Pixels: Determines image resolution and detail for accurate text extraction.
10. Variation: Measures noise level, influencing the clarity of text recognition.
11. Entropy: Indicates image complexity, affecting text extraction reliability.
12. Percentage Metrics: Offer insights into brightness distribution, aiding in text segmentation and extraction.
13. Angle: The skew angle of the document refers to the rotation angle required to align the document horizontally.
14. Segment Count: The number of segments refers to the count of identified regions containing text.

References

1. Butt, H.; Raza, M.R.; Ramzan, M.J.; Ali, M.J.; Haris, M. Attention-based CNN-RNN Arabic text recognition from natural scene images. *Forecasting* **2021**, *3*, 520–540. [CrossRef]
2. Bose, P.; Srinivasan, S.; Sleeman IV, W.C.; Palta, J.; Kapoor, R.; Ghosh, P. A survey on recent named entity recognition and relationship extraction techniques on clinical texts. *Appl. Sci.* **2021**, *11*, 8319. [CrossRef]
3. Chu, Q.; Chen, C.P.; Hu, H.; Wu, X.; Han, B. iHand: Hand Recognition-Based Text Input Method for Wearable Devices. *Computers* **2024**, *13*, 80. [CrossRef]
4. Jung, K.; Kim, K.I.; Jain, A.K. Text information extraction in images and video: A survey. *Pattern Recognit.* **2004**, *37*, 977–997. [CrossRef]
5. Zhang, H.; Zhao, K.; Song, Y.Z.; Guo, J. Text extraction from natural scene image: A survey. *Neurocomputing* **2013**, *122*, 310–323. [CrossRef]
6. Ye, Q.; Doermann, D. Text detection and recognition in imagery: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *37*, 1480–1500. [CrossRef]
7. Yin, X.C.; Zuo, Z.Y.; Tian, S.; Liu, C.L. Text detection, tracking and recognition in video: A comprehensive survey. *IEEE Trans. Image Process.* **2016**, *25*, 2752–2773. [CrossRef] [PubMed]
8. Zhao, M.; Li, S.; Kwok, J. Text detection in images using sparse representation with discriminative dictionaries. *Image Vis. Comput.* **2010**, *28*, 1590–1599. [CrossRef]
9. Neumann, L.; Matas, J. A method for text localization and recognition in real-world images. In Proceedings of the Computer Vision–ACCV 2010: 10th Asian Conference on Computer Vision, Queenstown, New Zealand, 8–12 November 2010; Revised Selected Papers, Part III 10; Springer: Berlin/Heidelberg, Germany, 2011; pp. 770–783.
10. Yi, C.; Tian, Y. Text extraction from scene images by character appearance and structure modeling. *Comput. Vis. Image Underst.* **2013**, *117*, 182–194. [CrossRef]
11. Gupta, N.; Banga, V. Image Segmentation for Text Extraction. In Proceedings of the 2nd International Conference on Electrical, Electronics and Civil Engineering (ICEECE'2012), Singapore, 28–29 April 2012; pp. 182–185.
12. Dang, N.C.; Moreno-García, M.N.; De la Prieta, F. Sentiment analysis based on deep learning: A comparative study. *Electronics* **2020**, *9*, 483. [CrossRef]
13. Dong, S.; Wang, P.; Abbas, K. A survey on deep learning and its applications. *Comput. Sci. Rev.* **2021**, *40*, 100379. [CrossRef]
14. Lu, H.; Ma, L.; Fu, X.; Liu, C.; Wang, Z.; Tang, M.; Li, N. Landslides information extraction using object-oriented image analysis paradigm based on deep learning and transfer learning. *Remote Sens.* **2020**, *12*, 752. [CrossRef]
15. Yang, Y.; Wu, Z.; Yang, Y.; Lian, S.; Guo, F.; Wang, Z. A survey of information extraction based on deep learning. *Appl. Sci.* **2022**, *12*, 9691. [CrossRef]
16. Yu, W.; Lu, N.; Qi, X.; Gong, P.; Xiao, R. PICK: Processing key information extraction from documents using improved graph learning-convolutional networks. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; pp. 4363–4370.
17. Revathi, A.; Modi, N.A. Comparative analysis of text extraction from color images using tesseract and opencv. In Proceedings of the 2021 8th International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, 17–19 March 2021; pp. 931–936.

18. Montejo-Ráez, A.; Jiménez-Zafra, S.M. Current approaches and applications in natural language processing. *Appl. Sci.* **2022**, *12*, 4859. [CrossRef]

19. Wang, D.; Su, J.; Yu, H. Feature extraction and analysis of natural language processing for deep learning English language. *IEEE Access* **2020**, *8*, 46335–46345. [CrossRef]

20. Olivetti, E.A.; Cole, J.M.; Kim, E.; Kononova, O.; Ceder, G.; Han, T.Y.J.; Hiszpanski, A.M. Data-driven materials research enabled by natural language processing and information extraction. *Appl. Phys. Rev.* **2020**, *7*, 041317. [CrossRef]

21. El-Komy, A.; Shahin, O.R.; Abd El-Aziz, R.M.; Taloba, A.I. Integration of computer vision and natural language processing in multimedia robotics application. *Inf. Sci.* **2022**, *7*, 765–775.

22. Laique, S.N.; Hayat, U.; Sarvepalli, S.; Vaughn, B.; Ibrahim, M.; McMichael, J.; Qaiser, K.N.; Burke, C.; Bhatt, A.; Rhodes, C.; et al. Application of optical character recognition with natural language processing for large-scale quality metric data extraction in colonoscopy reports. *Gastrointest. Endosc.* **2021**, *93*, 750–757. [CrossRef]

23. Wang, Y.; Wang, L.; Rastegar-Mojarad, M.; Moon, S.; Shen, F.; Afzal, N.; Liu, S.; Zeng, Y.; Mehrabi, S.; Sohn, S.; et al. Clinical information extraction applications: A literature review. *J. Biomed. Inform.* **2018**, *77*, 34–49. [CrossRef]

24. Viani, N.; Larizza, C.; Tibollo, V.; Napolitano, C.; Priori, S.G.; Bellazzi, R.; Sacchi, L. Information extraction from Italian medical reports: An ontology-driven approach. *Int. J. Med. Inform.* **2018**, *111*, 140–148. [CrossRef]

25. Hahn, U.; Oleynik, M. Medical information extraction in the age of deep learning. *Yearb. Med. Inform.* **2020**, *29*, 208–220. [CrossRef] [PubMed]

26. Dash, G.; Sharma, C.; Sharma, S. Sustainable marketing and the role of social media: An experimental study using natural language processing (NLP). *Sustainability* **2023**, *15*, 5443. [CrossRef]

27. Zhou, M.; Duan, N.; Liu, S.; Shum, H.Y. Progress in neural NLP: Modeling, learning, and reasoning. *Engineering* **2020**, *6*, 275–290. [CrossRef]

28. McMillan-Major, A.; Osei, S.; Rodriguez, J.D.; Ammanamanchi, P.S.; Gehrmann, S.; Jernite, Y. Reusable templates and guides for documenting datasets and models for natural language processing and generation: A case study of the HuggingFace and GEM data and model cards. *arXiv* **2021**, arXiv:2108.07374.

29. Zolotarev, O.; Solomentsev, Y.; Khakimova, A.; Charnine, M. Identification of semantic patterns in full-text documents using neural network methods. In Proceedings of the 29th International Conference on Computer Graphics and Vision, Graphicon-2019, Bryansk, Russia, 23–26 September 2019.

30. Zolotarev, O.; Research and development of linguo-statistical methods for forming a portrait of a subject area. In Proceedings of the International Conference "Computing for Physics and Technology—CPT2020", Moscow, Russia, 9–13 November 2020; Volume 11, pp. 305–313.

31. Kanev, A.I.; Savchenko, G.A.; Grishin, I.A.; Vasiliev, D.A.; Duma, E.M. Sentiment analysis of multilingual texts using machine learning methods. In Proceedings of the 2022 Conference of Russian Young Researchers in Electrical and Electronic Engineering (ElConRus), St. Petersburg, Russia, 25–28 January 2022; pp. 326–331.

32. Atnashev, T.; Ganeeva, V.; Kazakov, R.; Matyash, D.; Sonkin, M.; Voloshina, E.; Serikov, O.; Artemova, E. Razmecheno: Named Entity Recognition from Digital Archive of Diaries Prozhito. *arXiv* **2022**, arXiv:2201.09997.

33. da Silva, A.R.; Savić, D. Linguistic patterns and linguistic styles for requirements specification: Focus on data entities. *Appl. Sci.* **2021**, *11*, 4119. [CrossRef]

34. Bargshady, G.; Zhou, X.; Deo, R.C.; Soar, J.; Whittaker, F.; Wang, H. Enhanced deep learning algorithm development to detect pain intensity from facial expression images. *Expert Syst. Appl.* **2020**, *149*, 113305. [CrossRef]

35. Percha, B. Modern clinical text mining: A guide and review. *Annu. Rev. Biomed. Data Sci.* **2021**, *4*, 165–187. [CrossRef] [PubMed]

36. Graffelman, J.; De Leeuw, J. Improved approximation and visualization of the correlation matrix. *Am. Stat.* **2023**, *77*, 432–442. [CrossRef]

37. Sinha, P.; Crucilla, S.; Gandhi, T.; Rose, D.; Singh, A.; Ganesh, S.; Mathur, U.; Bex, P. Mechanisms underlying simultaneous brightness contrast: Early and innate. *Vis. Res.* **2020**, *173*, 41–49. [CrossRef]

38. Qiao, L.; Li, Z.; Cheng, Z.; Zhang, P.; Pu, S.; Niu, Y.; Ren, W.; Tan, W.; Wu, F. LGPMA: Complicated table structure recognition with local and global pyramid mask alignment. In Proceedings of the International Conference on Document Analysis and Recognition, Lausanne, Switzerland, 5–10 September 2021; Springer: Berlin/Heidelberg, Germany, 2021; pp. 99–114.

39. Singla, S. AI and IoT in healthcare. In *Internet of Things Use Cases for the Healthcare Industry*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 1–23.

40. Cao, L. Ai in finance: Challenges, techniques, and opportunities. *ACM Comput. Surv. (CSUR)* **2022**, *55*, 64 .

41. Zhai, X.; Chu, X.; Chai, C.S.; Jong, M.S.Y.; Istenic, A.; Spector, M.; Liu, J.B.; Yuan, J.; Li, Y. A Review of Artificial Intelligence (AI) in Education from 2010 to 2020. *Complexity* **2021**, *2021*, 8812542. [CrossRef]

42. Ranaldi, L.; Pucci, G. Knowing knowledge: Epistemological study of knowledge in transformers. *Appl. Sci.* **2023**, *13*, 677. [CrossRef]

43. Suster, D. A Mid-blue Logic. In *HUMAN RATIONALITY Festschrift for Nenad Smokrović*; Berčić, B., Golubović, A., Trobok, M., Eds.; University of Rijeka, Faculty of Humanities and Social Sciences: Rijeka, Croatia, 2022.

44. Paaß, G.; Konya, I. Machine learning for document structure recognition. In *Modeling, Learning, and Processing of Text Technological Data Structures*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 221–247.

45. Ghazal, T. Convolutional neural network based intelligent handwritten document recognition. *Comput. Mater. Contin.* **2022**, *70*, 4563–4581.
46. Akhter, M.P.; Jiangbin, Z.; Naqvi, I.R.; Abdelmajeed, M.; Mehmood, A.; Sadiq, M.T. Document-level text classification using single-layer multisize filters convolutional neural network. *IEEE Access* **2020**, *8*, 42689–42707. [CrossRef]
47. Lozić, E.; Štular, B. Documentation of archaeology-specific workflow for airborne LiDAR data processing. *Geosciences* **2021**, *11*, 26. [CrossRef]
48. Pizarro-Carmona, V.; Castano-Solis, S.; Cortes-Carmona, M.; Fraile-Ardanuy, J.; Jimenez-Bermejo, D. GA-based approach to optimize an equivalent electric circuit model of a Li-ion battery-pack. *Expert Syst. Appl.* **2021**, *172*, 114647. [CrossRef]
49. Rodrigues Mendes Ribeiro, R.; Dias Maciel, C. Bayesian Network Structural Learning Using Adaptive Genetic Algorithm with Varying Population Size. *Mach. Learn. Knowl. Extr.* **2023**, *5*, 1877–1887. [CrossRef]
50. Rezaeian, N.; Gurina, R.; Saltykova, O.A.; Hezla, L.; Nohurov, M.; Reza Kashyzadeh, K. Novel GA-Based DNN Architecture for Identifying the Failure Mode with High Accuracy and Analyzing Its Effects on the System. *Appl. Sci.* **2024**, *14*, 3354. [CrossRef]
51. Alblooshi, E.; Alblooshi, A.; Poon, K.; Ouali, A. A GA based approach for solving ring design telecommunication network. In Proceedings of the Genetic and Evolutionary Computation Conference Companion, Boston, MA, USA, 9–13 July 2022; pp. 415–418.
52. Shameem, M.; Nadeem, M.; Zamani, A.T. Genetic algorithm based probabilistic model for agile project success in global software development. *Appl. Soft Comput.* **2023**, *135*, 109998. [CrossRef]
53. Sengupta, S.; Basak, S.; Peters, R.A. Particle Swarm Optimization: A survey of historical and recent developments with hybridization perspectives. *Mach. Learn. Knowl. Extr.* **2018**, *1*, 157–191. [CrossRef]
54. Wang, K.; Li, X.; Gao, L.; Li, P.; Gupta, S.M. A genetic simulated annealing algorithm for parallel partial disassembly line balancing problem. *Appl. Soft Comput.* **2021**, *107*, 107404. [CrossRef]
55. Cherian, I.; Tamboli, A.I.; Pandey, A.; Manchanda, M.; Verma, G. A Comparative Study of Simulated Annealing and Ant Colony Optimization for Optimizing MRI-Based Alzheimer's Disease Classification. *Int. J. Intell. Syst. Appl. Eng.* **2024**, *12*, 464–475.
56. Sangaiah, A.K.; Hosseinabadi, A.A.R.; Shareh, M.B.; Bozorgi Rad, S.Y.; Zolfagharian, A.; Chilamkurti, N. IoT resource allocation and optimization based on heuristic algorithm. *Sensors* **2020**, *20*, 539. [CrossRef] [PubMed]
57. Park, A.; Hartzler, A.L.; Huh, J.; McDonald, D.W.; Pratt, W. Automatically detecting failures in natural language processing tools for online community text. *J. Med. Internet Res.* **2015**, *17*, e4612. [CrossRef] [PubMed]
58. Kim, S.H. Comparisons in End-to-End Pipeline Designs for Customized Document Information Extraction. Ph.D. Thesis, Massachusetts Institute of Technology, Cambridge, MA, USA , 2024.
59. Ben Tamou, A.; Benzinou, A.; Nasreddine, K. Live Fish Species Classification in Underwater Images by Using Convolutional Neural Networks Based on Incremental Learning with Knowledge Distillation Loss. *Mach. Learn. Knowl. Extr.* **2022**, *4*, 753–767. [CrossRef]
60. Coyac-Torres, J.E.; Sidorov, G.; Aguirre-Anaya, E.; Hernández-Oregón, G. Cyberattack Detection in Social Network Messages Based on Convolutional Neural Networks and NLP Techniques. *Mach. Learn. Knowl. Extr.* **2023**, *5*, 1132–1148. [CrossRef]
61. Mohammed, F.A.; Tune, K.K.; Assefa, B.G.; Jett, M.; Muhie, S. Medical Image Classifications Using Convolutional Neural Networks: A Survey of Current Methods and Statistical Modeling of the Literature. *Mach. Learn. Knowl. Extr.* **2024**, *6*, 699–735. [CrossRef]
62. Mayr, F.; Yovine, S.; Visca, R. Property Checking with Interpretable Error Characterization for Recurrent Neural Networks. *Mach. Learn. Knowl. Extr.* **2021**, *3*, 205–227. [CrossRef]
63. Yousef, M.; Hussain, K.F.; Mohammed, U.S. Accurate, data-efficient, unconstrained text recognition with convolutional neural networks. *Pattern Recognit.* **2020**, *108*, 107482. [CrossRef]
64. Lin, J.C.W.; Shao, Y.; Djenouri, Y.; Yun, U. ASRNN: A recurrent neural network with an attention model for sequence labeling. *Knowl.-Based Syst.* **2021**, *212*, 106548. [CrossRef]