applied sciences

MDPI

*Review*

# Advancing OCR Accuracy in Image-to-LaTeX Conversion—A Critical and Creative Exploration

**Everistus Zeluwa Orji [1,*], Ali Haydar [1], İbrahim Erşan [1] and Othmar Othmar Mwambe [2]**

[1] Department of Computer Engineering, Girne American University, Mersin-10, Karaman 99320, Turkey; ahaydar@gau.edu.tr (A.H.); ibrahimersan@gau.edu.tr (İ.E.)
[2] Computer Studies Department, Dar es Salaam Institute of Technology (DIT), Dar es Salaam P.O. Box 2958, Tanzania; othmar.mwambe@dit.ac.tz
* Correspondence: orjizeluwa@gmail.com

**Abstract:** This paper comprehensively assesses the application of active learning strategies to enhance natural language processing-based optical character recognition (OCR) models for image-to-LaTeX conversion. It addresses the existing limitations of OCR models and proposes innovative practices to strengthen their accuracy. Key components of this study include the augmentation of training data with LaTeX syntax constraints, the integration of active learning strategies, and the employment of active learning feedback loops. This paper first examines the current weaknesses of OCR models with a particular focus on symbol recognition, complex equation handling, and noise moderation. These limitations serve as a framework against which the subsequent research methodologies are assessed. Augmenting the training data with LaTeX syntax constraints is a crucial strategy for improving model precision. Incorporating symbol relationships, wherein contextual information is considered during recognition, further enriches the error correction. This paper critically examines the application of active learning strategies. The active learning feedback loop leads to progressive improvements in accuracy. This article underlines the importance of uncertainty and diversity sampling in sample selection, ensuring that the dynamic learning process remains efficient and effective. Appropriate evaluation metrics and ensemble techniques are used to improve the operational learning effectiveness of the OCR model. These techniques allow the model to adapt and perform more effectively in diverse application domains, further extending its utility.

**Keywords:** optical character recognition (OCR); LaTeX; active learning strategies; image-to-LaTeX conversion; natural language processing (NLP)



**Citation:** Orji, E.Z.; Haydar, A.; Erşan, İ.; Mwambe, O.O. Advancing OCR Accuracy in Image-to-LaTeX Conversion—A Critical and Creative Exploration. *Appl. Sci.* **2023**, *13*, 12503. https://doi.org/10.3390/app132212503

Academic Editor: Douglas O'Shaughnessy

Received: 13 September 2023
Revised: 1 November 2023
Accepted: 7 November 2023
Published: 20 November 2023



## 1. Introduction

The digital age has transformed how we interact with written content, with optical character recognition (OCR) technology serving as a linchpin in this transformation [1]. OCR enables the conversion of printed or handwritten text into machine-readable formats, thus ushering in an era of enhanced accessibility and utility for textual data. Even though the accurate recognition and conversion of mathematical expressions into LaTeX format is still a challenge that looms large, it is within this complex and critical arena that we find the motivation and contributions of this study. The rationale for undertaking this research is underpinned by a profound recognition of the crucial role played by mathematical expression recognition within the broader OCR landscape. Mathematical notation, characterized by its intricate symbols and complex structures, has long been a vexing challenge for OCR systems [2]. The precise recognition and correct conversion of mathematical expressions require understanding the symbols themselves and a deep grasp of the semantics, syntax, and intricate relationships interweaving these symbols [1]. However, prevailing OCR methods, while formidable, often fall short of capturing these subtleties, resulting in conversions that do not meet the stringent accuracy requirements [3]. The crux of the challenge lies

in the visual complexity of mathematical symbols, where symbols that bear a striking resemblance can possess distinct semantic meanings [4]. The perennial noise or distortion in input images adds a layer of complexity, directly impeding the OCR system's ability to recognize and interpret mathematical expressions accurately.

Recognizing these multifaceted challenges propels our quest for innovative solutions at the intersection of OCR and natural language processing (NLP) to enhance the accuracy of mathematical expression recognition and conversion. Hence, this extensive review study is aimed at exploring various existing natural language processing (NLP) techniques that attempt to enhance OCR accuracy in image-to-LaTeX conversions. This study also analyzes the limitations of existing approaches and recommends future directions. In turn, this study exposes existing research gaps and paves the way for innovative NLP integration techniques in OCR. In order to meet these research goals, this review study has gone through several stages (see Figure 1). These include a thorough literature review that sets out the problem statement and tries to answer the research question about how NLP techniques can be added to OCR to make it more accurate when converting images to LaTeX, an analytical screening of the techniques introduced by various research articles, and the recommendation of future directions.

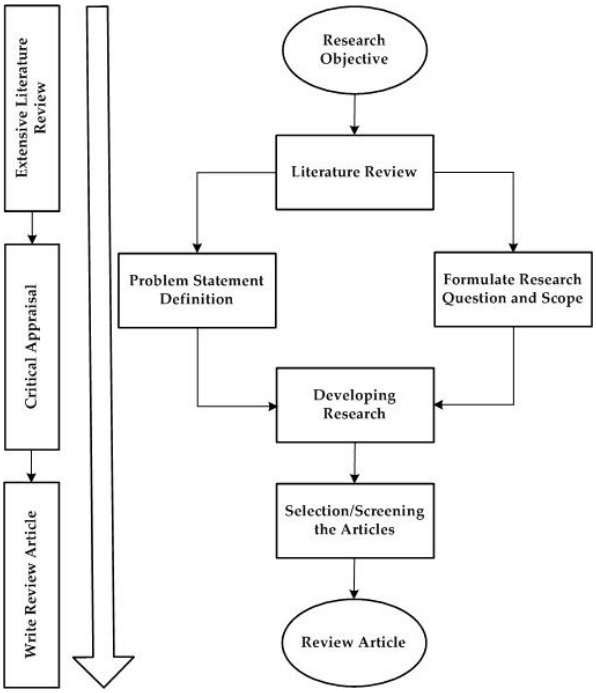


**Figure 1.** Research method.

The remainder of this paper is structured as follows: the background information of this study and the related works will be stated in Section 2; the deep learning strategies for OCR in the image-to-LaTeX conversion will be stated in Section 3; the preprocessing techniques for image enhancement will be presented in Section 4; the limitations of current OCR models will be presented in Section 5; the augmenting OCR training data with LaTeX Syntax constraints will be presented in Section 6; binarization and thresholding techniques will be presented in Section 7; leveraging symbol relationships for OCR error correction will be described in Section 8; post-processing techniques for error correction in OCR for image-to-LaTeX conversion will be presented in Section 9; post-processing strategies to leverage the redundancy inherent in mathematical notation will be presented in Section 10; active learning strategies for incorporating the OCR model will be described in Section 11; evaluation metrics for OCR accuracy in image-to-LaTeX conversion will be presented in Section 12; and finally, Section 13 will conclude this study and provide recommendations for future directions.

## 2. Background Information

The previous OCR approaches have struggled with handling the complexities and intricacies of mathematical notation, causing suboptimal conversions. Mathematical scholars have turned to natural language processing (NLP) techniques to respond to these challenges to enhance OCR accuracy in image-to-LaTeX conversion. Integrating NLP strategies into OCR models can potentially improve mathematical expression recognition and conversion by leveraging semantic information and including linguistic context. Applying deep learning architectures like recurrent neural networks, convolutional neural networks, and transformer models, can enable OCR systems to capture subtle patterns and relationships within mathematical expressions [5].

Advancements in retraining techniques, like bidirectional encoder representations from transformers, can also capture the contextual embedding that helps accurately recognize mathematical symbols. This paper's main objective is to critically examine the challenges associated with OCR accuracy in the image-to-LaTeX conversion and propose creative solutions to enhance the performance of OCR models. This report explores innovative techniques that leverage NLP methods to address the limitations of current OCR systems. The main limitation of current OCR models in accurately recognizing and converting mathematical expressions is handling the complex equations characterized by several symbols and intricate structural arrangements [2]. Identifying and interpreting such equations is critical for accurate conversion to LaTeX format. Various mathematical symbols also tend to be challenging due to their visual similarity, making it difficult for OCR models to differentiate between similar-looking symbols accurately [4]. Another challenge arises from noise or distortion in the input images that negatively affects OCR performance (see Table 1). OCR models can better interpret signs based on their surrounding context and infer their intended meaning by encoding contextual information [6]. For instance, understanding whether a symbol represents an operator, a variable, or a function is essential for accurate conversion to LaTeX.

**Table 1.** Summary of challenges and techniques in OCR approaches to image-to-LaTeX conversion.

| OCR Approach | Performance | Challenges and Techniques |
|---|---|---|
| NLP Integration [7] | Potential | Complex equations and intricate structural arrangements.<br>Visual similarity of mathematical symbols.<br>Noise or distortion in input images.<br>Leveraging semantic information and linguistic context through NLP techniques. |
| Data Augmentation [8] | Improved | Ensuring adherence to LaTeX syntax rules.<br>Reducing generation of syntactically incorrect LaTeX code.<br>Incorporating LaTeX syntax constraints during data augmentation. |
| Dependency Capture [9] | Enhanced | Capturing dependencies and relationships between symbols (subscripts, superscripts, fraction components).<br>Improving accuracy by correcting recognition errors and ensuring the integrity of the converted LaTeX representation. |
| Active Learning [10] | Efficient | Mitigating dependency on large labeled datasets.<br>Intelligently selecting informative and challenging examples.<br>Involving human annotators in training through active learning strategies. (query-by-committee, uncertainty sampling, adaptive sampling). |

Incorporating LaTeX syntax constraints during data augmentation guides OCR models to learn more accurate and compliant conversions, ensuring that the output adheres to LaTeX syntax rules. Studies can reduce the likelihood of generating syntactically incorrect LaTeX code during conversion by enforcing these constraints [5]. The constraint-based augmentation strategy improves accuracy and reliability in OCR outputs [11]. Additionally, mathematical expressions often exhibit dependencies and relationships between symbols, such as subscripts, superscripts, or fraction components. We can enhance the accuracy of OCR outputs by correcting recognition errors and ensuring the integrity of the converted

LaTeX representation by capturing these dependencies and incorporating them into the OCR model. Traditional OCR approaches require large labeled datasets for training, which are expensive, time-consuming, and costly to create [12]. Active learning addresses this challenge by intelligently selecting samples for manual annotation to mitigate the dependency on large labeled datasets. The OCR model can focus on learning from the most informative and challenging examples, leading to more efficient and effective model improvement by actively involving human annotators in training [4]. Active learning strategies like query-by-committee, uncertainty sampling, and adaptive sampling, can be used to select samples that maximize the model's learning potential.

## 3. Deep Learning Strategies for OCR in Image-to-LaTeX Conversion

Deep learning strategies have reformed the field of optical character recognition (OCR) and enhanced the accuracy in image-to-LaTeX conversion. Leveraging neural network architectures like recurrent neural networks and convolutional neural networks enables researchers to tackle the complex challenges of character recognition and equation parsing [13,14]. The learning models rely on large-scale annotated datasets to learn and generalize. However, developing high-quality datasets that include various styles, fonts, and mathematical symbols poses significant challenges. It is critical to address data collection challenges to ensure representative and unbiased training sets. Studies have tried creating benchmark datasets specific to image-to-LaTeX conversion, like the CROHME dataset containing handwritten mathematical expressions [15–18]. The datasets enhance the evaluation and training of OCR models and serve as a foundation for advancing the field [1]. CNNs and RNNs may struggle with out-of-domain and rare symbols encountered in image-to-LaTeX conversion, leading to low accuracy and errors in the converted LaTeX result.

Addressing this challenge requires domain-specific knowledge and designing models that can handle the intricacies of mathematical notation. For instance, studies have examined integrating mathematical grammar rules into OCR models to enable the recognition process and enhance accuracy [19]. These models can achieve more reliable conversions by incorporating mathematical semantics and structure. Understanding how the models predict is vital for identifying and rectifying OCR issues. However, the inherent challenge of deep learning architectures deters their interpretability [12]. Studies have developed methods for visualizing the attention and feature activations in CNNs and RNNs. However, further advancement is essential to ensure transparent and reliable OCR systems. Explainable systems like saliency analysis and attention maps can provide insight into the decision-making process of OCR models and help identify potential sources of errors [20]. The computational necessities of deep learning models also pose a challenge since training and deploying complex neural network skill sets require serious computational resources. This challenge hinders the scalability and accessibility of OCR systems in resource-constrained environments.

Assessing techniques for model compression, hardware acceleration, and granting them enough skill, may mitigate these difficulties and make OCR solutions more practical. Studies have recommended lightweight OCR models that attain comparable accuracy to larger ones since they require fewer computational resources to enhance their deployment on low-power devices [12]. OCR errors at the character recognition stage may also develop during the conversion process, leading to substantial inaccuracies in the final LaTeX output. Creating error correction techniques and post-processing mechanisms is vital for mitigating the impact of OCR errors and ensuring high-quality conversions. Recent studies have examined applying language models and contextual information to the enhance error correction in OCR outputs [21]. Leveraging contextual clues and syntactic analysis enables these strategies to identify and rectify OCR errors, thus fostering the accuracy of the converted LaTeX representations.

## 4. Preprocessing Techniques for Image Enhancement

Preprocessing techniques for image enhancement in OCR for image-to-LaTeX conversion are crucial for improving the accuracy of the recognition process [22,23]. These techniques address challenges related to image quality, noise, contrast, skew, and multimodal features. Recent advancements in this field have shown promising results, but critical aspects still need to be considered. One aspect is noise reduction and image denoising techniques. The noise in input images can significantly impact OCR accuracy [4,24]. Researchers have proposed various denoising algorithms, such as median filtering, Gaussian filtering, and wavelet-based methods, to reduce noise and artifacts. Additionally, recent studies have introduced advanced denoising algorithms based on deep learning approaches, leveraging convolutional autosencoders and generative adversarial networks (GANs) [25]. These methods have demonstrated improved OCR accuracy by effectively suppressing noise patterns and preserving the legibility of characters and symbols.

Another crucial preprocessing step is contrast enhancement. Enhancing image contrast can significantly improve the readability of characters and symbols, especially in low-quality or poorly illuminated images (see Figure 2). Histogram equalization techniques, such as adaptive histogram equalization (AHE) and contrast-limited adaptive histogram equalization (CLAHE), have been widely used. Recent research has explored integrating deep learning models, such as U-Net and Pix2Pix networks, for adaptive contrast enhancement [25]. These approaches have demonstrated their effectiveness in handling varying illumination conditions and improving OCR performance. Binarization and thresholding techniques are also critical in OCR preprocessing. Binarization converts grayscale or color images into binary representations, separating foreground characters from the background [25]. Various thresholding techniques, including global thresholding, local adaptive thresholding, and hybrid methods, have been proposed to address different image characteristics and challenges. Recent advancements have introduced deep learning-based binarization methods that utilize convolutional neural networks to learn optimal thresholding strategies. These approaches have shown promising results in handling complex backgrounds and improving the segmentation of characters, leading to improved OCR accuracy.

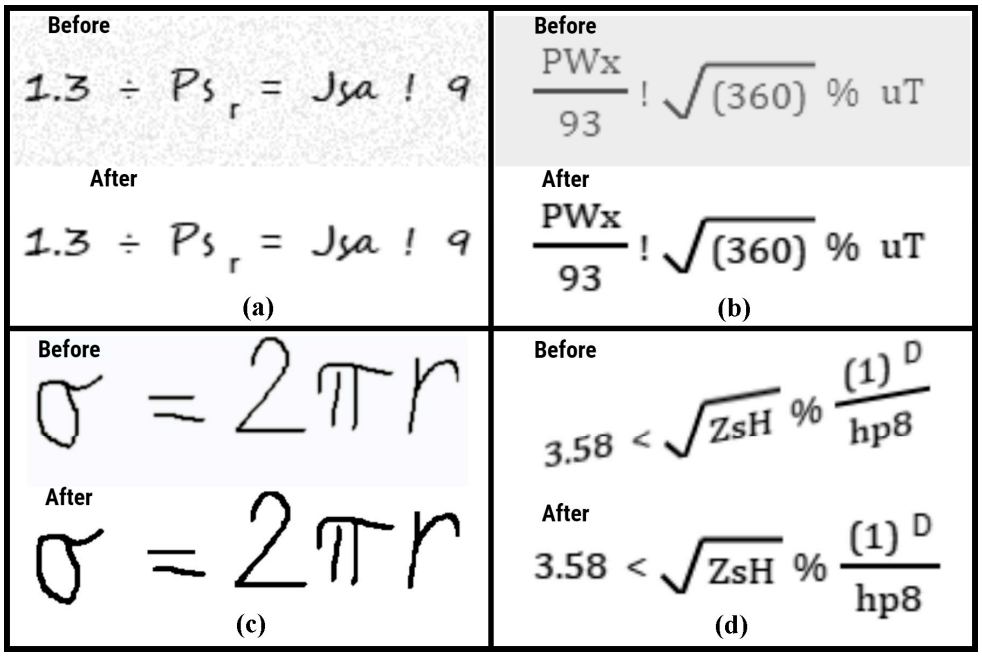


**Figure 2.** Visual comparison of image preprocessing techniques (noise reduction (**a**), contrast enhancement (**b**), binarization (**c**), and skew correction (**d**)) for OCR.

Skew detection and correction techniques are essential for aligning images and ensuring accurate character recognition. Skewed or rotated images can negatively impact OCR accuracy. Recent research has explored the use of deep learning models, such as convolutional neural networks and recurrent neural networks, for automatic skew detection and correction [25]. These models leverage learned features and geometric transformations to estimate and rectify image skew. These techniques improve OCR accuracy by effectively aligning the images, mainly when dealing with skewed documents. Multimodal fusion and feature enhancement techniques have also gained attention in OCR preprocessing [1]. OCR models can benefit from complementary features during preprocessing by fusing multiple modalities, such as color, texture, and shape information. Recent studies have investigated the fusion of these modalities to enhance the discriminative power of OCR models. Furthermore, attention mechanisms and contextual information have been explored to guide feature enhancement [25–27]. These approaches enable OCR systems to focus on informative regions while suppressing noise or irrelevant details, ultimately improving recognition accuracy.

While recent advancements in preprocessing techniques for image enhancement have shown promising results, there are still challenges to be addressed. Finding the right balance between noise suppression and the preservation of fine details is crucial. Adapting to varying image quality and the efficient handling of input image types, such as handwritten or scanned documents, also require further research. Moreover, the computational complexity of some deep learning-based techniques may limit their practicality, especially in resource-constrained environments. Future research should focus on developing robust preprocessing methods that are adaptive, efficient, and capable of handling diverse real-world scenarios to enhance OCR accuracy in image-to-LaTeX conversion [21].

## 5. Analyzing the Limitations of Current OCR Models

Optical character recognition (OCR) has improved the digitization of documents by enhancing the conversion of printed and longhand text into machine-readable formats. However, the accurate recognition and conversion of mathematical expressions from images to the LaTeX format remain challenging. The main limitation with current OCR models is handling complex equations accurately. Mathematical expressions involve variables, many symbols, operators, and nested structures, making them inherently difficult to interpret and convert accurately. OCR models should comprehend the hierarchical relationships between symbols and the intended mathematical operations to produce faithful LaTeX representations [4]. Complicated equations challenge OCR models to capture the exact arrangement of symbols and matrices and recognize fractions, subscripts, superscripts, and parentheses. Mathematical expressions include advanced concepts like Greek symbols, integrals, and summations, that further complicate recognition. Handling complex equations enables OCR models to provide reliable and accurate conversions to the LaTeX format [2].

OCR models also face challenges in accurately recognizing various mathematical symbols. Mathematical notation includes multiple characters, including numerals, operators, alphabets, Greek letters, and mathematical functions [28]. Most of these symbols show visual similarities, making it difficult for OCR models to differentiate between similar-looking characters accurately [1]. Differentiating between the symbol “0” and the letter “o” or distinguishing between the variable “x” and the multiplication operator “×” is a challenge for OCR models. Moreover, recognizing and differentiating between similar-looking symbols accurately, like “cos” and “sin” or “α” and “a”, requires OCR models to possess a robust symbol recognition capability that can handle variations in font sizes, styles, and orientations [5]. To overcome this limitation, OCR models may benefit from incorporating contextual information and leveraging the semantic relationships between symbols. OCR models can also make more informed decisions regarding symbol recognition and improve the accuracy of the conversion process by considering the surrounding context and the syntax of mathematical expressions [2].

Distortion in the input images also presents an excellent challenge for OCR models. Images captured in real-world scenarios are characterized by types of noise, such as pixelation, blurring, lighting, and artefacts introduced during the scanning process. Documents may have longhand annotations, erasures, or smudges that further complicate the recognition process. These distortions affect the clarity and legibility of mathematical expressions and cause errors in recognition and subsequent LaTeX conversion [25]. OCR models must be robust and resilient to warping to ensure accurate recognition and conversion. OCR models can benefit from preprocessing techniques for image enhancement and noise reduction. These techniques involve denoising, filtering, and contrast adjustment to improve the legibility of the input images before OCR processing [29]. Integrating noise-robust recognition algorithms and data augmentation techniques that simulate various types of noise can also improve the OCR model's ability to effectively handle noisy or distorted images. Mathematical expressions provide intricate correlations between symbols like subscripts, superscripts, and fraction components. OCR models can correct recognition errors and maintain the structural integrity of the converted LaTeX representation by modelling these relationships [30].

OCR models can be improved using advanced techniques that capture the hierarchical structure and semantic relationships within mathematical expressions to overcome their limitations when handling complex equations. OCR systems can understand the intricate arrangements of symbols and capture the nuances of mathematical notation by incorporating deep learning models such as convolutional neural networks and recurrent neural networks. These models can learn to recognize and interpret matrices, fractions, subscripts, and nested parentheses more accurately, improving conversion quality. Identifying various mathematical symbols requires OCR models to have comprehensive symbol recognition capabilities [4]. The old OCR models struggle with differentiating similar-looking characters, leading to conversion errors. Leveraging semantic information and contextual embedding will enable the models to understand symbols better based on their surrounding context and syntactic patterns. Contextual information can help differentiate between visually similar characters and improve symbol recognition accuracy. Addressing noise and distortion challenges requires noise-robust recognition algorithms and robust preprocessing techniques. OCR models can be trained using augmented datasets that simulate various types of noise to make them more resilient to real-world image imperfections. Combining strong preprocessing and noise-robust recognition algorithms can allow the models to handle challenging image conditions better and produce accurate conversions.

Incorporating domain-specific knowledge into OCR models can improve the models' performance. Mathematical expressions adhere to mathematical conversion and syntax rules. OCR models can learn to generate LaTeX outputs, conforming to syntactic rules by integrating LaTeX syntax constraints during training. The constraint-based data augmentation strategy enables the model to provide valid and compliant LaTeX code and mitigates the likelihood of producing syntactically incorrect conversions [1]. Active learning strategies can be employed to iteratively improve the model's performance and enhance OCR accuracy further. Traditional OCR approaches often rely on large labeled datasets for training, which is time-consuming and expensive. Active learning allows the model to actively select informative samples for manual annotation, reducing its dependency on extensive labeled datasets [29]. Active learning helps improve the model's performance with fewer labeled examples, resulting in more efficient and effective training by selecting challenging or uncertain samples for the model.

Evaluation metrics such as recall, precision, and F1 score can be applied to assess the performance of OCR models in symbol recognition. These metrics quantify the OCR model's completeness, accuracy, and overall performance in recognizing mathematical symbols [25]. Iterative improvements can be made to the OCR model based on the results. The process may involve fine-tuning the deep learning architectures, refining contextual information utilization, and adjusting training data augmentation strategies. By continuously evaluating and refining the OCR model, recognition accuracy can be enhanced,

leading to more reliable and precise image-to-LaTeX conversions. The visual similarities, contextual complexities, and font variations are obstacles to symbol recognition. However, utilizing semantic relationships, incorporating contextual information, augmenting training data, and incorporating domain-specific knowledge can enable OCR models to overcome these challenges. Improving symbol recognition accuracy in OCR models enables a more accurate and reliable conversion of mathematical expressions from images to LaTeX format. This advancement has significant implications for various fields, including academic research, scientific publishing, and document digitization. Enabling the efficient representation of mathematical content can help OCR models contribute to disseminating scientific knowledge and enhance the accessibility of mathematical information [31]. Continuously advancing and refining recognition capabilities can allow the models to better serve the needs of researchers, professionals, and educators who rely on accurate and efficient image-to-LaTeX conversion.

The semantic relationships between symbols provide additional indications for accurate recognition. Mathematical expressions show relationships such as fractions or superscript components [30]. OCR models can correct recognition errors and maintain the structural integrity of the converted LaTeX representation by modeling these relationships. For instance, recognizing a fraction requires identifying the denominator and numerator symbols and their relative positions. OCR models can enhance symbol recognition accuracy by leveraging these semantic relationships [25]. Transformer models have shown promising results in symbol recognition tasks. Convolutional neural networks are effective in capturing local visuals. OCR models can leverage their strengths to improve symbol recognition accuracy by combining these architectures. OCR models can benefit from augmented training datasets encompassing variations in symbol appearance, font styles, sizes, and orientations [1]. Data augmentation techniques like noise injections and random rotations can help the model learn to handle variations commonly encountered in real-world scenarios (see Table 2). Augmenting the training data with various symbol instances enables the OCR model to generalize and recognize symbols accurately. Domain-specific knowledge can enhance symbol recognition in OCR models [29]. OCR models can be trained to identify the appropriate usage of symbols in mathematical expressions, such as differentiating between using "$\pi$" as a constant or as a variable.

**Table 2.** Summary of challenges and techniques in optical character recognition (OCR) for the conversion of mathematical expressions.

| Challenges | Techniques |
| --- | --- |
| Handling complex equations [25]. | Incorporating deep learning models (e.g., convolutional neural networks, recurrent neural networks) to capture hierarchical structures and semantic relationships within mathematical expressions. |
| Recognizing various mathematical symbols [32]. | Leveraging semantic information and contextual embedding, integrating strong symbol recognition capabilities. |
| Dealing with noise and distortion in input images [33]. | Employing noise-robust recognition algorithms and robust preprocessing techniques, training OCR models using augmented datasets. |
| Incorporating domain-specific knowledge [34]. | Integrating LaTeX syntax constraints during training, utilizing active learning strategies, refining contextual information utilization. |
| Evaluating OCR model performance [35,36]. | Applying evaluation metrics (recall, precision, F1 score), refining OCR models through fine-tuning, adjusting training data augmentation strategies. |
| Enhancing symbol recognition accuracy [37]. | Utilizing semantic relationships, combining transformer models and convolutional neural networks, augmenting training datasets with variations in symbol appearance. |
| Handling variations in symbol usage and appearance [38]. | Augmenting training data with variations in symbol instances, incorporating domain-specific knowledge to identify appropriate symbol usage. |

## 6. Augmenting OCR Training Data with LaTeX Syntax Constraints

Augmenting OCR training data with LaTeX syntax constraints is a critical approach that improves OCR models' performances in image-to-LaTeX conversion. OCR models can generate LaTeX representations that adhere to the correct semantics, structure, and syntax of mathematical expressions by including the concept of LaTeX syntax rules and applying them during the training process. The main advantage of augmenting OCR training data with LaTeX syntax constraints is the improvement in conversion accuracy. LaTeX syntax provides a well-defined and standardized framework for representing mathematical notation. The models can learn to recognize and correct common errors made during the OCR and conversion processes by training OCR models with augmented data that includes the correct LaTeX syntax [31]. For example, OCR models can be trained to locate missing LaTeX delimiters like brackets, parentheses, or curly braces and rectify them accordingly. This ability to correct errors leads to the enhanced accuracy and validity of the converted LaTeX code. Augmenting OCR training data with LaTeX syntax constraints also enables the preservation of structural integrity. LaTeX syntax rules define the hierarchical correlation between mathematical symbols, subexpressions, and operators. Incorporating this knowledge during training can help OCR models understand mathematical expressions' structure and organization [31]. The information allows the models to generate LaTeX code that accurately represents the structural integrity of the original mathematical content. For example, OCR models can learn to correctly handle superscripts, fractions, subscripts, and nested parentheses.

Augmenting OCR training data with LaTeX syntax constraints also enables a semantic understanding of mathematical expressions. LaTeX syntax provides semantic concepts that give information about the meaning and interpretation of the mathematical content. For instance, applying specific LaTeX commands or environments indicates the type of mathematical concept represented, like equations, matrices, or mathematical functions [30]. Training OCR models with augmented data containing these semantic concepts enables the model to understand mathematical content better [39]. This understanding allows them to generate more meaningful LaTeX code that reflects the intended semantics of the original mathematical expressions. Integrating LaTeX syntax constraints into OCR training data improves consistency in the LaTeX representation of mathematical expressions. Following the LaTeX syntax rules ensures that the generated LaTeX code is consistent and coherent when dealing with mathematical notations with multiple hierarchy levels [29]. Augmenting OCR training data with LaTeX syntax constraints enables OCR models to produce LaTeX code that adheres to mathematical expressions' expected structure and semantics, leading to a more coherent and readable output.

OCR models trained with augmented data incorporating LaTeX syntax constraints also show improved compatibility with existing LaTeX tools, workflows, and libraries. OCR models produce outputs seamlessly integrated with the LaTeX environment by generating LaTeX code conforming to the syntax constraints. This compatibility enhances further manipulation, processing, and rendering of the converted LaTeX code to enable users to leverage existing LaTeX tools for tasks such as typesetting, rendering to PDF or other formats, and educational materials [40]. Augmenting OCR training data with LaTeX syntax constraints also enables a reduction in post-processing work. OCR models can minimize the need for extensive manual correction and post-processing of the converted content by generating LaTeX code that adheres to syntax rules [41]. This saves time and effort for users who rely on accurate and reliable image-to-LaTeX conversion. The reduction in post-processing requirements allows users to focus on other essential activities such as validation, content analysis, or further correction of the converted LaTeX code. LaTeX is broadly applied in academic and scientific domains for typesetting mathematical content. OCR models produce outputs that seamlessly integrate into existing LaTeX workflows, tools, and publishing pipelines by generating LaTeX code that conforms to the constraints of syntax [30]. The compatibility improves the usability and practicality of the converted LaTeX code and enables users to leverage the full potential of LaTeX for further processing.

In summary, augmenting OCR training data with LaTeX syntax constraints provides advantages for enhancing the performance and reliability of OCR models in image-to-LaTeX conversion. It enables consistency, accuracy, and semantic understanding while mitigating post-processing work and ensuring compatibility with existing LaTeX workflows and tools [42]. OCR models can generate LaTeX representations that capture mathematical expressions' semantics, structure, and syntax by incorporating the knowledge of LaTeX syntax rules.

## 7. Binarization and Thresholding Techniques

The performance of learning models relies on large-scale annotated datasets. However, developing high-quality datasets with diverse styles and mathematical symbols poses a severe challenge due to the scarcity of datasets designed for image-to-LaTeX conversion. Scholars have tried to gather datasets representing the difficulties encountered in the conversion process [43]. The Competition on Recognition of Handwritten Mathematical Expressions (CHROME) dataset incorporates handwritten mathematical expressions that enable the evaluation and comparison of OCR models in the context of image-to-LaTeX conversion [44]. The MathML and LaTeX dataset (MALL) is also concerned with mathematical expressions expressed in MathML and LaTeX formats, enhancing the training and assessment of OCR systems for this specific task [45,46]. These benchmark datasets create a foundation for strengthening the field by enabling standardized assessment and fostering the development of more accurate OCR models [19]. The quality and representativeness of the training data are vital since any form of bias presented in the training data may affect the accuracy of OCR systems.

Training data mainly composed of a specific font or style makes it difficult for the OCR model to recognize symbols and characters from other styles and fonts accurately. Addressing such challenges requires keen data collection and annotation plans to ensure a balanced and diverse representation of styles, fonts, and mathematical symbols. Collaborative strategies among academic institutions and scholars can play a critical role in addressing this challenge by ensuring diversity, sharing datasets, and fostering a more comprehensive understanding of the intricacies of mathematical notation [20]. In the context of training data, the scarcity of data for rare symbols is also a critical area to be addressed. Deep learning models perform better on frequently occurring characters and styles in the training data. Scholars have examined synthetic data generation and augmentation techniques [12]. Data augmentation includes applying diverse transformations like scaling, rotation, and adding noise to augment the training data and expose the model to broad variations. Synthetic data generation is the development of artificial images with rare notations to supplement the training data. These techniques enhance OCR models' generalization capability and improve their accuracy when working with less common symbols.

The training data's diversity and size are other essential elements that affect OCR accuracy. The learning models need large-scale training data to learn the images' underlying variations and patterns. Insufficient training data leads to overfitting, in which the model fails to generalize well to unseen examples. The diversity of the training data is also critical to ensuring generalization to different styles, fonts, and writing styles [42]. Gathering a comprehensive and diverse dataset is a non-trivial task, since it requires the consideration of variations in handwriting, mathematical domains, and notation styles. Studies should expand and diversify the available training data to improve the accuracy and reliability of OCR systems for image-to-LaTeX conversion [47]. Training data biased toward specific cultural or regional preferences lead to inaccurate OCR results. For instance, OCR models trained on datasets that mainly encompass Western mathematical notation may struggle when faced with symbols or notes used in non-Western languages and mathematical systems [31]. It is essential to incorporate diverse cultural perspectives and collaborations with experts from various regions to ensure the comprehensive coverage of mathematical notations and symbols.

## 8. Leveraging Symbol Relationships for OCR Error Correction

Leveraging symbol relationships for OCR error correction enhances the accuracy and reliability of image-to-LaTeX conversion. Mathematical notation contains interconnected symbols and operators that convey specific relationships and meanings. OCR models can detect and rectify errors during recognition and conversion by understanding and analyzing these symbol relationships [48]. OCR systems face errors when dealing with complex mathematical equations and symbols. OCR models can identify and correct these errors by considering the context and relationships between characters. For instance, examining the relationships between adjacent symbols can enable OCR models to detect and rectify errors like missing symbols [31]. This strategy helps ensure that the converted LaTeX code accurately represents the original mathematical expression [49].

Symbol relationships also play an important role in error correction related to the positioning of superscripts and subscripts. OCR errors may lead to incorrectly positioned subscripts and superscripts, affecting the meaning of mathematical expressions [29]. OCR models can analyze symbols' relative heights and alignments to determine the correct positioning of subscripts and superscripts by leveraging symbol relationships. This assessment enables the models to mitigate errors and accurately represent mathematical notation. Complex equations with brackets or nested parentheses also challenge OCR systems [13]. Errors may occur when opening and closing symbols, leading to missing and imbalanced delimiters. OCR models can analyze the relationships between opening and closing symbols to detect and rectify such errors by leveraging symbol relationships. For example, when a closing parenthesis is missing, the OCR model can identify the corresponding opening parenthesis and insert the missing closing symbol. This approach ensures the correct representation of the structural integrity of mathematical expressions.

Leveraging symbol relationships also enhances the correction of errors related to the misinterpretation of mathematical operators. OCR errors may occur when operators interpret incorrectly, leading to incorrect mathematical representations. OCR models can analyze the context and identify the correct operator based on its relationship with adjacent symbols by considering the relationships between symbols. This enables the models to correct errors and ensure the accurate representation of mathematical operations. It is crucial to effectively train OCR models with data about symbol relationships to leverage symbol relationships for OCR error correction. OCR models learn to recognize individual symbols and understand their relationships during training [48]. This contextual understanding enables the models to assess symbol sequences, apply error correction strategies, and identify potential errors based on symbol relationships. OCR models develop a deeper understanding of mathematical notation and can make informed decisions about error correction by incorporating symbol relationships into the training process [50].

Contextually understanding symbol relationships allows OCR models to make informed decisions about error correction, leading to more reliable and accurate conversions. OCR models ensure the structural integrity and semantic accuracy of the converted LaTeX code by rectifying errors related to subscripts, superscripts, delimiters, and operators. The primary technique used in leveraging symbol relationships is the application of context windows [51]. OCR models analyze a window of symbols surrounding the symbol in question to determine its correct identity and position. The models can make informed decisions about error correction by considering the neighboring symbols and their relationships. For instance, when a symbol is interpreted as a division operator instead of a fraction bar, the OCR model can examine the symbols before and after the fraction bar to identify the correct interpretation based on the context [13]. Utilizing information about the mathematical domain and the relationships between the symbols also enables OCR models to enhance error correction accuracy [52].

Machine learning algorithms can be trained to explain symbol relationships for error correction. These models learn to recognize symbols and the correlation between them by analyzing large annotated datasets. These models can capture the dependencies and patterns in mathematical notation by incorporating symbol relationships into the training

process to make accurate predictions and corrections during the OCR and conversion processes. Graph-based approaches can be used to leverage symbol relationships in mathematical expressions, in the form of graphs, to identify and correct errors [52]. For instance, a disconnected node in the graph can infer the missing symbol and restore the right relationship between symbols. The spatial arrangement of symbols within a mathematical expression may also provide important information about their relationships [13]. OCR models can examine symbols' relative positions, alignments, and distances to infer their roles and relationships. For instance, when two symbols are vertically aligned, they are likely to be related as a numerator and denominator in a fraction [43]. Leveraging symbol relationships can also be complemented with statistical concepts like probabilistic models. The models estimate the likelihood of certain symbol relationships based on the statistical properties of mathematical notation. OCR models can make informed decisions and prioritize the most likely corrections by incorporating statistical information into the error correction process.

In summary, leveraging symbol relationships for OCR error correction requires using semantic information, context-based machine-learning algorithms, statistical techniques, graph-based approaches, and spatial relationships. The courses enable OCR models to examine the relationship between symbols and make accurate predictions and corrections during recognition and conversion [53]. The tools enable OCR systems to achieve higher accuracy, improve the structural integrity of mathematical expressions, and ensure the semantic fidelity of the converted LaTeX code.

## 9. Post-Processing Techniques for Error Correction in OCR for Image-to-LaTeX Conversion

Post-processing techniques improve the accuracy of Optical Character Recognition (OCR) systems for image-to-LaTeX conversion by addressing errors that occur during the recognition process. OCR models are not infallible, and mistakes propagate and increase throughout the conversion process [11]. The creation of robust error correction strategies is essential to ensuring high-quality conversions. One critical aspect of post-processing is error detection. Various error detection methodologies include statistical analysis, linguistic analysis, and pattern matching. A study compared the OCR output with statistical models to identify discrepancies [22]. Statistical techniques can identify potential errors based on their deviation from the expected patterns by analyzing the frequency and distribution of symbols. Pattern-matching strategies apply regular expressions to identify the mistakes in the OCR output. The designs can capture inconsistencies in the OCR results, like misrecognized symbols. Linguistic analysis leverages language models and grammar rules to identify semantic errors. It can identify mistakes that violate grammatical or mathematical rules by analyzing the OCR output in the context of the surrounding text or equations [28]. By combining these approaches, error detection algorithms can identify potential errors and flag them for further correction.

Error correction techniques come into play upon detecting errors to rectify the OCR and enhance the accuracy of the final LaTeX output. Language models provide contextual information and semantic understanding to correct mistakes in OCR outputs [43]. The models consider equations and the surrounding words to identify and rectify errors. If the OCR output contains a misspelled and unrecognized word, the language model provides alternative words based on the context to improve the accuracy of the converted LaTeX representation [54]. Language models can give valuable suggestions for error correction by leveraging the statistical properties of language and the context in which OCR errors occur. The integration of contextual information is also used to correct mistakes [55]. Contextual analysis includes analyzing the relationships between equations, symbols, and mathematical expressions to identify and correct errors. OCR errors that disrupt the overall coherence of the mathematical expressions are detected and rectified by considering the syntactic and structural context. If an OCR error results in an equation that violates mathematical rules, the contextual analysis identifies the discrepancy and proposes corrections that maintain the integrity of the equation.

## 10. Post-Processing Strategies Leverage the Redundancy Inherent in Mathematical Notation

Mathematical expressions have various representations that convey similar meanings. The redundancy can be in the form of alternative notes, equivalent forms of equations, or mathematical transformations [56]. Assessing redundancy enables error correction algorithms to find the most probable corrected version of the OCR output and improve the accuracy of the converted LaTeX representation [28]. Error correction techniques ensure the correctness and consistency of the converted LaTeX representation by utilizing mathematical equivalences and transformations. Leveraging external knowledge bases and resources improves error correction in OCR for image-to-LaTeX conversion [25]. These resources include domain-specific databases, mathematical ontologies, and mathematical libraries. Errors can be identified and corrected based on the known correct representations by comparing the OCR output against these resources. Domain-specific rules can be included in the error correction process to foster the accuracy and consistency of the converted LaTeX representation. For instance, if the OCR output contains a mathematical symbol that is incompatible with the mathematical domain or context, the error correction mechanism can propose appropriate replacements based on the domain-specific rules [39]. Error correction techniques can improve the accuracy and reliability of the OCR system for image-to-LaTeX conversion by integrating external knowledge and domain-specific information.

The accuracy of error correction techniques depends on the quality and accuracy of the OCR output. The correction process becomes more challenging if the OCR system has many errors. Therefore, it is vital to continuously enhance and correct the underlying OCR algorithms to mitigate recognition errors [22]. Improvements in preprocessing techniques like noise reduction, image enhancement, and segmentation lead to better OCR results and consequently enhance the effectiveness of error correction in image-to-LaTeX conversion. The complexity of mathematical notation poses challenges for error correction for OCR in image-to-LaTeX conversion [47]. Mathematical expressions involve intricate symbols and mathematical notations specific to different domains (see Table 2). Enhancing the accurate recognition and modification of these elements requires specialized algorithms and techniques tailored to the complexities of mathematical notation [4]. Developing domain-specific models and algorithms and collaborating with mathematicians and domain skill sets leads to more accurate error correction in mathematical OCR.

## 11. Incorporating Active Learning Strategies for OCR Model Improvement

The main advantage of incorporating active learning is its effectiveness in the annotation process. The old OCR training required annotating a large dataset with ground truth labels, which is expensive and time-consuming [57]. Active learning reduces these challenges by prioritizing the most informative samples for annotation [2]. Active learning reduces the annotation effort while ensuring effective model training by selectively choosing samples for which the model is uncertain or likely to make errors in. This efficient use of annotation resources saves time and reduces the costs associated with the training phase. It can also improve OCR model performance by allowing the model to learn from its mistakes and refine its understanding of symbol recognition and conversion. This iterative process allows the model to concentrate on critical areas for improvement, leading to more accurate and reliable image-to-LaTeX conversions [39]. The model becomes more robust and capable of handling various symbols, font styles, and mathematical structures encountered in real-world scenarios by actively targeting challenging samples for annotation [43].

The feedback loop created between the model and the annotation process leads to a progressive cycle of learning and refinement. As the OCR model encounters new data and challenging samples during the image-to-LaTeX conversion, it flags those samples for annotation. Experts then annotate the samples, and the newly labeled data updates the model [50]. The updated model has been equipped with additional knowledge that is then applied to the conversion process to enhance performance. This iterative cycle enables the OCR model to adapt and improve continuously. Active learning also enhances the

generalization and adaptability of OCR models [43]. The models learn to handle various symbol variations, mathematical structures, font styles, and noise patterns by incorporating diverse samples through strategies like diversity sampling. This generalization capability enables the model to handle different handwritings, unique mathematical notations, and variations in expression formats [58]. It also fosters collaboration between the OCR system and human experts. The ability of the model to flag challenging samples for annotation enables human experts to provide their expertise and domain knowledge [59]. Experts contribute to improving the OCR model's performance by annotating these samples. This collaboration enhances the quality of the training dataset since human experts can validate and correct errors made by the model.

A critical aspect to consider is selecting the active learning query strategy. Various query strategies determine how the model chooses samples for annotation. Common query strategies include query-by-committee, uncertainty sampling, and diversity sampling [60]. Uncertainty sampling selects samples with low confidence scores, indicating the model's uncertainty about their correct labels. Diversity sampling targets a diverse range of samples to ensure comprehensive training [61]. Query-by-committee includes training multiple models with slightly different initializations or architectures and selecting samples on which these models disagree, thus targeting areas of uncertainty. Deciding on an appropriate query strategy depends on the OCR system's specific requirements and the nature of the data.

It is also critical to consider the balance between exploration and exploitation. Examination involves selecting samples that the model has not seen before, enabling it to learn from diverse examples, while exploitation focuses on picking pieces expected to provide the most significant improvement in the model's performance. Striking the right balance between the two ensures that the OCR model continues to learn and improve while maximizing its accuracy on challenging samples [11]. Active learning may also benefit from including ensemble techniques. The techniques combine multiple OCR models, each trained on different subsets of the training data or with different architectures, to make predictions. Ensemble models often provide more robust and accurate predictions by aggregating the knowledge and insights from multiple models. In the context of active learning, ensemble models can be utilized to improve the reliability of sample selection [43]. The operational learning strategy can make more informed decisions about which samples to annotate, further enhancing the model's performance by considering the agreement or disagreement among ensemble members on the uncertainty of samples.

The choice of evaluation metrics is important when incorporating active learning into OCR models. The original metrics, like error or accuracy rates, might not be sufficient to capture the nuances of OCR performance [62]. Metrics that consider the complexity of mathematical expressions can provide a more comprehensive assessment of the OCR system's performance. Applying appropriate evaluation metrics can optimize the active learning process by focusing on challenging samples that directly impact the overall quality of the image-to-LaTeX conversion. Domain adaptation techniques can also be utilized to improve the efficiency of active learning in OCR [63]. Due to domain differences, OCR models trained on synthetic data might struggle to perform well on real-world documents. The model can be fine-tuned on a small amount of real-world data, making it more capable of handling the specific challenges and variations present in real-world OCR scenarios by leveraging domain adaptation methods [64]. Incorporating domain adaptation into the active learning pipeline ensures that the samples selected for annotation align with the target domain, resulting in improved performance and accuracy. Active learning for OCR is dynamic, and various techniques and strategies are continuously explored [59]. Research efforts focus on developing more sophisticated and efficient sample selection approaches, investigating the integration of active learning with other methods, and leveraging advanced machine learning algorithms. These progressive developments aim to enhance the capabilities of OCR models further and optimize the dynamic learning process for image-to-LaTeX conversion.

In summary, incorporating active learning strategies into OCR models for image-to-LaTeX conversion has numerous advantages [65]. The efficiency of the annotation process is enhanced by selectively choosing the best samples for annotation, while also mitigating the annotation effort and related costs. The performance of the model is enhanced through iterative learning and refinement. The continuous feedback loop ensures that the model adapts and improves over time, keeping pace with dynamic challenges [62]. Active learning also enhances the model's adaptability and generalization by incorporating diverse samples. The correlation between the OCR system and human experts enriches the training dataset and improves the model's performance.

## 12. Evaluation Metrics for OCR Accuracy in Image-to-LaTeX Conversion

Evaluation metrics are important for examining the accuracy and performance of optical character recognition (OCR) systems in the context of image-to-LaTeX conversion. The metrics provide quantitative measures that enable studies to compare different OCR algorithms, track field progress, and identify areas for improvement [66]. Analyzing the accuracy of OCR in the image-to-LaTeX conversion presents unique challenges due to the complex nature of mathematical notation and the need for accurate representation in the LaTeX format [44]. The main aspect to evaluating OCR accuracy is comparing the OCR output with ground truth references [15]. The latter represents the correct LaTeX representation of the mathematical expressions contained in the images. Comparing the OCR output against these references enhances the calculation of metrics that measure the similarity between the OCR result and the ground truth [49]. Various strategies can be utilized for the comparison, such as semantic analysis, string-matching algorithms, and structural similarity measures. These techniques enable the quantification of the accuracy of the OCR system in terms of symbol recognition, overall fidelity, and equation structure [29].

Symbol-level evaluation metrics examine OCR systems' accuracy in recognizing individual symbols in mathematical expressions. The metrics include recall, precision, and F1 score, commonly used in pattern recognition activities. These metrics enable studies to assess the performance of OCR systems in accurately identifying and recognizing mathematical symbols [67]. Equation-level evaluation metrics examine the accuracy of OCR systems in capturing the structure and syntax of mathematical expressions [68]. The matrices include an arrangement of symbols and adherence to mathematical rules. The most commonly applied metric is equation-level accuracy, which measures the proportion of correctly recognized structured equations. The structural similarity metric quantifies the similarity between the OCR output and the ground truth regarding the hierarchical structure and relationships between symbols [69]. These metrics provide information about the OCR system's ability to preserve the structural integrity of mathematical expressions during the conversion process.

Semantic evaluation metrics also examine OCR systems' accuracy in capturing the semantic meaning of mathematical expressions. Mathematical notation enables various equivalent representations, and preserving the semantic equivalence is important for correct conversion to LaTeX [54]. Semantic evaluation metrics include the similarity between the semantic models of the OCR output and the ground truth. Strategies like semantic matching, parsing, and embedding may be utilized to measure semantic similarity and assess the OCR system's accuracy in capturing the intended meaning of mathematical expressions. Domain-specific evaluation metrics are critical for determining OCR accuracy in specialized mathematical domains [25]. Various mathematical disciplines may have specific symbols, notations, or conventions that must be correctly recognized and represented in LaTeX [70]. Evaluating OCR systems in these domains requires domain-specific evaluation metrics that capture the complexities and intricacies of the notation. Collaborating with mathematicians, domain experts, and educators is vital for defining and developing these metrics and ensuring that OCR systems meet the requirements of specific mathematical domains [43]. The evaluation of OCR accuracy in image-to-LaTeX conversion should take into account the efficiency and computational complexity of the OCR systems. Large-scale document

processing requires OCR systems to provide accurate results within acceptable time frames. Evaluation metrics that include processing speed, scalability, and resource utilization can comprehensively assess the OCR systems' performance in practical scenarios.

## 13. Conclusions, Limitations, and Recommendations

This study examines optical character recognition (OCR) for image-to-LaTeX conversion, mainly focusing on the transformative potential of active learning strategies. It presents a holistic approach to advancing OCR accuracy, addressing the limitations of current OCR models, introducing innovative techniques, and emphasizing the critical role of context-aware processing, active learning, and domain adaptation in achieving this goal. The limitations of existing OCR models are multifaceted, encompassing challenges related to recognizing mathematical symbols, handling complex equations, and effectively managing noise in the input data [71]. OCR models have struggled to cope with the intricacies of mathematical notation and the diverse typographical conventions associated with LaTeX documents. Symbol recognition has been a persistent challenge due to variations in writing styles and the complex interplay of symbols within equations.

This research first introduces the concept of augmenting training data with LaTeX syntax constraints to address these limitations. This innovative approach entails constraining the OCR model's predictions to adhere to LaTeX syntax rules during training. Integrating LaTeX-specific controls enhances the model's understanding of mathematical expressions, enabling it to discern structure and semantics accurately. Consequently, the OCR system produces LaTeX representations that align with LaTeX syntax, resulting in higher accuracy and reliability in image-to-LaTeX conversions. Symbol relationships within mathematical expressions also assume critical significance in our exploration [2]. OCR models gain the ability to rectify errors and enhance the fidelity of LaTeX conversions by considering the contextual information and interdependencies of the symbols. This context-aware processing approach represents a crucial step towards overcoming the challenges associated with symbol recognition and understanding. It explains the importance of context in OCR, providing a roadmap for further advancements in symbol recognition, interpretation, and conversion accuracy. Active learning introduces a dynamic element to the OCR process by helping the model to selectively choose informative samples for annotation. This strategic selection enhances the model's performance by focusing on challenging areas and refining its understanding of symbol recognition and conversion. The active learning feedback loop, combined with ensemble techniques and appropriate evaluation metrics, creates a progressive learning and refinement cycle, allowing OCR models to adapt and improve over time. This iterative process significantly enhances accuracy and reliability, making it an invaluable tool in the OCR toolkit.

This study also emphasizes the importance of uncertainty and diversity sampling in active learning. These strategies ensure that the dynamic learning process remains efficient and effective, carefully balancing exploration and exploitation. Ensemble models strengthen OCR predictions' robustness and accuracy through their ability to aggregate knowledge from multiple models. Ensemble-based sample selection plays a pivotal role in the effectiveness of active learning strategies. Domain adaptation techniques emerge as a crucial aspect of our research since they allow OCR models trained on synthetic data to be fine-tuned on real-world data, aligning them to the specific challenges and variations encountered in practical OCR scenarios. Methods such as unsupervised or semi-supervised learning are instrumental in enhancing the transferability and effectiveness of active learning strategies when faced with real-world complexities [13].

This study offers a comprehensive framework for advancing OCR accuracy in image-to-LaTeX conversion. Through active learning and domain adaptation, this research paves the way for more accurate and versatile OCR systems by tackling the limitations of current OCR models, proposing innovative methodologies, and highlighting the importance of context-aware processing. The implications of this work extend far and wide, from improving scientific documentation and mathematical education to enhancing accessibility for

visually impaired individuals. This paper underscores the significance of incorporating techniques that allow OCR models trained on synthetic data to perform well on real-world documents to address the domain adaptation challenge. The OCR models can be fine-tuned on a small amount of real-world data, setting them in alignment with the specific challenges and variations encountered in practical OCR scenarios, by leveraging domain adaptation methods such as unsupervised or semi-supervised learning. Including domain adaptation considerations enhances the transferability and effectiveness of active learning strategies in real-world applications.

It is crucial to recognize various limitations that have influenced the course and application of this study, despite its significant impacts. Firstly, the dynamic technological landscape poses a constant challenge. This study has been conducted within a rapidly evolving field where new algorithms, hardware, and software emerge daily. Continuous updates and adaptations are needed to keep pace with the ever-evolving landscape of technology and maintain the cutting-edge applicability of our methods. Secondly, this study might not cover every possible detail and difficulty OCR professionals face. The domain of mathematical expressions and LaTeX texts is broad and complex, and although our approaches show promise, they might not cover every particular situation. Changes in LaTeX conventions and variations in mathematical notations may present new difficulties that require specific considerations.

Furthermore, even with the best attempts to use domain adaptation techniques to close the gap between synthetic and real-world data, real-world OCR applications might still pose unique challenges that require additional improvements and customized methods. The variety in real-world documents, including differences in writing styles and the subtleties of symbol usage, can provide difficulties beyond our investigation's purview. The subjective character of human interpretation and assessment may result in variations in the perceived accuracy of OCR outcomes. As human decisions and subjectivity can affect the impact of our innovations, the human element in accuracy assessment suggests that judgments of the system's performance may differ. These limitations give future researchers an important direction and a firm base to build on as they improve and advance our work.

**Author Contributions:** E.Z.O., A.H., İ.E. and O.O.M. have equally contributed to the conceptualization, methodology, writing—original draft preparation, writing—review and editing, and writing and revision of this paper, while supervision was performed by A.H. and İ.E. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Data are contained within the article.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Drobac, S.; Lindén, K. Optical character recognition with neural networks and post-correction with finite state methods. *Int. J. Doc. Anal. Recognit.* **2020**, *23*, 279–295. [CrossRef]
2. Garkal, A.; Pal, A.; Singh, K.P. HMER-Image to LaTeX: A Variational Dropout Approach. In Proceedings of the 2021 5th Conference on Information and Communication Technology (CICT), Kurnool, India, 10–12 December 2021. [CrossRef]
3. Deng, Y.; Yu, Y.; Yao, J.; Sun, C. An Attention Based Image to Latex Markup Decoder. In Proceedings of the 2017 Chinese Automation Congress (CAC), Jinan, China, 20–22 October 2017; pp. 7199–7203. [CrossRef]
4. Kayal, P.; Anand, M.; Desai, H.; Singh, M. Tables to LaTeX: Structure and content extraction from scientific tables. *Int. J. Doc. Anal. Recognit.* **2022**, *26*, 121–130. [CrossRef]
5. Bitterman, D.S.; Goldner, E.; Finan, S.; Harris, D.; Durbin, E.B.; Hochheiser, H.; Warner, J.L.; Mak, R.H.; Miller, T.; Savova, G.K. An End-to-End Natural Language Processing System for Automatically Extracting Radiation Therapy Events From Clinical Texts. *Int. J. Radiat. Oncol.* **2023**, *117*, 262–273. [CrossRef]
6. Heo, T.S.; Kim, Y.S.; Choi, J.M.; Jeong, Y.S.; Seo, S.Y.; Lee, J.H.; Jeon, J.P.; Kim, C. Prediction of Stroke Outcome Using Natural Language Processing-Based Machine Learning of Radiology Report of Brain MRI. *J. Pers. Med.* **2020**, *10*, 286. [CrossRef] [PubMed]

7. Rokde, C.N.; Kshirsagar, D.M. NLP challenges for machine translation from english to indian languages. *Int. J. Comput. Sci. Inform.* **2020**, *4*, 5. [CrossRef]
8. Wei, J.; Wang, Q.; Song, X.; Zhao, Z. The Status and Challenges of Image Data Augmentation Algorithms. *J. Phys. Conf. Ser.* **2023**, *2456*, 012041. [CrossRef]
9. Ritz, F.; Phan, T.; Sedlmeier, A.; Altmann, P.; Wieghardt, J.; Schmid, R.; Sauer, H.; Klein, C.; Linnhoff-Popien, C.; Gabor, T. Capturing Dependencies Within Machine Learning via a Formal Process Model. *Lect. Notes Comput. Sci.* **2022**, *13703*, 249–265. [CrossRef]
10. Vodovozov, V.; Raud, Z.; Petlenkov, E. Challenges of Active Learning in a View of Integrated Engineering Education. *Educ. Sci.* **2021**, *11*, 43. [CrossRef]
11. Jin, J.A. The Evolution of Visual Spectacle: A Virtual-Reality Exhibition at the Charles B. Wang Center. *Ars Orient.* **2020**, *50*, 20220203. [CrossRef]
12. Jaderberg, M.; Simonyan, K.; Vedaldi, A.; Zisserman, A. Deep Structured Output Learning for Unconstrained Text Recognition. *arXiv* **2014**, arXiv:1412.5903.
13. Yang, J.; Drake, T.; Damianou, A.; Maarek, Y. Leveraging Crowdsourcing Data for Deep Active Learning an Application: Learning Intents in alexa. In Proceedings of the 2018 World Wide Web Conference, Lyon, France, 23–27 April 2018; pp. 23–32. [CrossRef]
14. Najam, R.; Faizullah, S. Analysis of Recent Deep Learning Techniques for Arabic Handwritten-Text OCR and Post-OCR Correction. *Appl. Sci.* **2023**, *13*, 7568. [CrossRef]
15. Beyerer, J.; Puente León, F.; Frese, C.; Beyerer, J.; Puente León, F.; Frese, C. Preprocessing and Image Enhancement. In *Machine Vision: Automated Visual Inspection: Theory, Practice and Applications*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 465–519. [CrossRef]
16. Mouchère, H.; Viard-Gaudin, C.; Kim, D.H.; Kim, J.H.; Garain, U. CROHME2011: Competition on Recognition of Online Handwritten Mathematical Expressions. Available online: https://hal.science/hal-00615216/file/CROHME_CRC511.pdf (accessed on 22 July 2023).
17. Mouchère, H.; Viard-Gaudin, C.; Kim, D.H.; Kim, J.H.; Garain, U. ICFHR 2012-Competition on Recognition of On-line Mathematical Expressions (CROHME 2012). Available online: http://www.isical.ac.in/~crohme (accessed on 22 July 2023).
18. Mouchère, H.; Viard-Gaudin, C.; Zanibbi, R.; Garain, U.; Kim, D.H.; Kim, J.H. ICDAR 2013 CROHME: Third International Competition on Recognition of Online Handwritten Mathematical Expressions. 2013. Available online: www.isical.ac.in/ (accessed on 22 July 2023).
19. Deng, Y.; Kanervisto, A.; Ling, J.; Rush, A.M. Image-to-Markup Generation with Coarse-to-Fine Attention. In Proceedings of the International Conference on Machine Learning, New York, NY, USA, 19–24 June 2016.
20. Sivaramakrishnan, A.; Kumar, M.V. Pre-Processing and Image Enhancement Techniques. *IJARCCE* **2020**, *9*, 107–113. [CrossRef]
21. Wang, Z.; Liu, J.C. PDF2LaTeX: A Deep Learning System to Convert Mathematical Documents from PDF to LaTeX. In Proceedings of the ACM Symposium on Document Engineering 2020, New York, NY, USA, 29 September–1 October 2020. [CrossRef]
22. Saddami, K.; Munadi, K.; Away, Y.; Arnia, F. Effective and fast binarization method for combined degradation on ancient documents. *Heliyon* **2019**, *5*, e02613. [CrossRef] [PubMed]
23. Lim, C.C.; Ling, A.H.W.; Chong, Y.F.; Mashor, M.Y.; Alshantti, K.; Aziz, M.E. Comparative Analysis of Image Processing Techniques for Enhanced MRI Image Quality: 3D Reconstruction and Segmentation Using 3D U-Net Architecture. *Diagnostics* **2023**, *13*, 2377. [CrossRef]
24. Shopon, M.; Diptu, N.A.; Mohammed, N. End-to-End Optical Character Recognition Using Sythetic Dataset Generator for Noisy Conditions. In Proceedings of the International Joint Conference on Computational Intelligence: IJCCI 2019, Dhaka, Bangladesh, 25–26 October 2019; pp. 515–527. [CrossRef]
25. Zhou, M.; Cai, M.; Li, G.; Li, M. An End-to-End Formula Recognition Method Integrated Attention Mechanism. *Mathematics* **2022**, *11*, 177. [CrossRef]
26. Huang, Z.; Ma, Y.; Wang, R.; Li, W.; Dai, Y. A Model for EEG-Based Emotion Recognition: CNN-Bi-LSTM with Attention Mechanism. *Electronics* **2023**, *12*, 3188. [CrossRef]
27. Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. *arXiv* **2017**, arXiv:1706.03762.
28. Hino, H. Active Learning: Problem Settings and Recent Developments. 2020. Available online: https://arxiv.org/abs/2012.04225v2 (accessed on 13 June 2023).
29. Liu, Y.; Li, Z.; Li, H.; Yu, W.; Huang, M.; Peng, D.; Liu, M.; Chen, M.; Li, C.; Liu, C.-L.; et al. On the Hidden Mystery of OCR in Large Multimodal Models. 2023. Available online: https://arxiv.org/abs/2305.07895v3 (accessed on 13 June 2023).
30. Wang, X.; Liu, Q.; Gui, T.; Zhang, Q.; Zou, Y.; Zhou, X.; Ye, J.; Zhang, Y.; Zheng, R.; Pang, Z.; et al. TextFlint: Unified Multilingual Robustness Evaluation Toolkit for Natural Language Processing. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations, Online, 1–6 August 2021; pp. 347–355. [CrossRef]
31. Zhang, J.; Sang, J.; Xu, K.; Wu, S.; Zhao, X.; Sun, Y.; Hu, Y.; Yu, J. Robust CAPTCHAs Towards Malicious OCR. *IEEE Trans. Multimedia* **2021**, *23*, 2575–2587. [CrossRef]
32. Kukreja, V. Sakshi Machine learning models for mathematical symbol recognition: A stem to stern literature analysis. *Multimedia Tools Appl.* **2022**, *81*, 28651–28687. [CrossRef]

33. Ogwok, D.; Ehlers, E.M. Detecting, Contextualizing and Computing Basic Mathematical Equations from Noisy Images using Machine Learning. In Proceedings of the 2020 3rd International Conference on Computational Intelligence and Intelligent Systems, Tokyo, Japan, 13–15 November 2020; pp. 8–14. [CrossRef]
34. Lu, M.; Fang, Y.; Yan, F.; Li, M. Incorporating Domain Knowledge into Natural Language Inference on Clinical Texts. *IEEE Access* **2019**, *7*, 57623–57632. [CrossRef]
35. Karpinski, R.; Lohani, D.; Belaid, A. Metrics for Complete Evaluation of OCR Performance. 2018. Available online: https://inria.hal.science/hal-01981731 (accessed on 27 October 2023).
36. Neudecker, C.; Baierer, K.; Gerber, M.; Clausner, C.; Antonacopoulos, A.; Pletschacher, S. A Survey of OCR Evaluation Tools and Metrics. In Proceedings of the 6th International Workshop on Historical Document Imaging and Processing, Lausanne, Switzerland, 5–6 September 2021; pp. 13–18. [CrossRef]
37. Bin, O.K.; Hooi, Y.K.; Kadir, S.J.A.; Fujita, H.; Rosli, L.H. Enhanced Symbol Recognition based on Advanced Data Augmentation for Engineering Diagrams. *Int. J. Adv. Comput. Sci. Appl.* **2022**, *13*, 537–546. [CrossRef]
38. Patil, S.; Varadarajan, V.; Mahadevkar, S.; Athawade, R.; Maheshwari, L.; Kumbhare, S.; Garg, Y.; Dharrao, D.; Kamat, P.; Kotecha, K. Enhancing Optical Character Recognition on Images with Mixed Text Using Semantic Segmentation. *J. Sens. Actuator Networks* **2022**, *11*, 63. [CrossRef]
39. Tang, L.A.; Korona-Bailey, J.; Zaras, D.; Roberts, A.; Mukhopadhyay, S.; Espy, S.; Walsh, C.G. Using Natural Language Processing to Predict Fatal Drug Overdose from Autopsy Narrative Text: Algorithm Development and Validation Study. *JMIR Public Health Surveill* **2023**, *9*, e45246. [CrossRef] [PubMed]
40. Bilbeisi, G.; Ahmed, S.; Majumdar, R. DeepEquaL: Deep Learning Based Mathematical Equation to Latex Generation. In Proceedings of the Neural Information Processing: 27th International Conference, ICONIP 2020, Bangkok, Thailand, 18–22 November 2020; Volume 1333, pp. 324–332. [CrossRef]
41. Kaluarachchi, T.; Wickramasinghe, M. A systematic literature review on automatic website generation. *J. Comput. Lang.* **2023**, *75*, 101202. [CrossRef]
42. Maharana, K.; Mondal, S.; Nemade, B. A review: Data pre-processing and data augmentation techniques. *Glob. Transit. Proc.* **2022**, *3*, 91–99. [CrossRef]
43. Springmann, U.; Fink, F.; Schulz, K.U. Automatic Quality Evaluation and (Semi-) Automatic Improvement of OCR Models for Historical Printings. 2016. Available online: https://arxiv.org/abs/1606.05157v2 (accessed on 13 June 2023).
44. Shidaganti, G.; Salil, S.; Anand, P.; Jadhav, V. Robotic Process Automation with AI and OCR to Improve Business Process: Review. In Proceedings of the 2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 4–6 August 2021; pp. 1612–1618. [CrossRef]
45. Scharpf, P.; Schubotz, M.; Cohl, H.S.; Breitinger, C.; Gipp, B. Discovery and Recognition of Formula Concepts using Machine Learning. 2023. Available online: https://arxiv.org/abs/2303.01994v2 (accessed on 23 July 2023).
46. Gipp, B.; Greiner-Petter, A.; Schubotz, M.; Meuschke, N. Methods and Tools to Advance the Retrieval of Mathematical Knowledge from Digital Libraries for Search-, Recommendation-, and Assistance-Systems. *arXiv* **2023**, arXiv:2305.07335.
47. Pandey, S.; Pandey, S.K.; Miller, L. Measuring Innovativeness of Public Organizations: Using Natural Language Processing Techniques in Computer-Aided Textual Analysis. *Int. Public Manag. J.* **2016**, *20*, 78–107. [CrossRef]
48. Wang, J.; Sun, Y.; Wang, S. Image to Latex with DenseNet Encoder and Joint Attention. *Procedia Comput. Sci.* **2019**, *147*, 374–380. [CrossRef]
49. Chu, J.S.V.; Pyo, B.; Parth, V.; Hussein, A.; Wang, P. Key–Value Pair Identification from Tables Using Multimodal Learning. *Int. J. Pattern Recognit. Artif. Intell.* **2023**, *37*, 2352009. [CrossRef]
50. Hirlekar, V.V.; Kumar, A. Natural Language Processing based Online Fake News Detection Challenges—A Detailed Review. In Proceedings of the 2020 5th International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 10–12 June 2020; pp. 748–754. [CrossRef]
51. Borovikov, E. A Survey of Modern Optical Character Recognition Techniques. 2014. Available online: https://arxiv.org/abs/1412.4183v1 (accessed on 20 July 2023).
52. Sandnes, F.E. Lost in OCR-Translation: Pixel-based Text Reflow to the Rescue: Magnification of Archival Raster Image Documents in the Browser without Horizontal Scrolling. In Proceedings of the 15th International Conference on PErvasive Technologies Related to Assistive Environments, New York, NY, USA, 29 June–1 July 2022; pp. 500–506. [CrossRef]
53. Shruthi, J.; Swamy, S. A prior case study of natural language processing on different domain. *Int. J. Electr. Comput. Eng.* **2020**, *10*, 4928–4936. [CrossRef]
54. Crema, C.; Attardi, G.; Sartiano, D.; Redolfi, A. Natural language processing in clinical neuroscience and psychiatry: A review. *Front. Psychiatry* **2022**, *13*, 946387. [CrossRef] [PubMed]
55. Mehta, N.; Braun, P.X.; Gendelman, I.; Alibhai, A.Y.; Arya, M.; Duker, J.S.; Waheed, N.K. Repeatability of binarization thresholding methods for optical coherence tomography angiography image quantification. *Sci. Rep.* **2020**, *10*, 15368. [CrossRef] [PubMed]
56. Zhang, Z.; Zhang, Z.; Di Caprio, F.; Gu, G.X. Machine learning for accelerating the design process of double-double composite structures. *Compos. Struct.* **2022**, *285*, 115233. [CrossRef]
57. Li, M.; Zhao, P.; Zhang, Y.; Niu, S.; Wu, Q.; Tan, M. Structure-Aware Mathematical Expression Recognition with Sequence-Level Modeling. In Proceedings of the 29th ACM International Conference on Multimedia, New York, NY, USA, 20–24 October 2021; pp. 5038–5046. [CrossRef]

58. Dalal, J.; Daiya, S. Image Processing Based Optical Character Recognition Using Matlab. *Int. J. Eng. Sci. Res. Technol.* **2018**, *30*, 406–411. [CrossRef]
59. Edwards, K.M. Accelerating the Design Process Through Natural Language Processing-based Idea Filtering. 2022. Available online: https://dspace.mit.edu/handle/1721.1/147338 (accessed on 20 July 2023).
60. Jiang, K.; Lu, X. Natural Language Processing and Its Applications in Machine Translation: A Diachronic Review. In Proceedings of the 2020 IEEE 3rd International Conference of Safe Production and Informatization (IICSPI), Chongqing, China, 28–30 November 2020; pp. 210–214. [CrossRef]
61. Ling, X.; Gao, M.; Wang, D. Intelligent Document Processing Based on RPA and Machine Learning. In Proceedings of the 2020 Chinese Automation Congress (CAC), Shanghai, China, 6–8 November 2020; pp. 1349–1353. [CrossRef]
62. Wu, J.-W.; Yin, F.; Zhang, Y.-M.; Zhang, X.-Y.; Liu, C.-L. Image-to-markup generation via paired adversarial learning. *Lect. Notes Comput. Sci.* **2019**, *11051*, 18–34. [CrossRef]
63. Moon, N.N.; Salehin, I.; Parvin, M.; Hasan, M.; Talha, I.M.; Debnath, S.C.; Nur, F.N.; Saifuzzaman, M. Natural language processing based advanced method of unnecessary video detection. *Int. J. Electr. Comput. Eng.* **2021**, *11*, 5411–5419. [CrossRef]
64. Leaman, R.; Khare, R.; Lu, Z. Challenges in clinical natural language processing for automated disorder normalization. *J. Biomed. Inform.* **2015**, *57*, 28–37. [CrossRef]
65. Dong, L.-F.; Liu, H.-C.; Zhang, X.-M. Synthetic Data Generation and Shuffled Multi-Round Training Based Offline Handwritten Mathematical Expression Recognition. *J. Comput. Sci. Technol.* **2022**, *37*, 1427–1443. [CrossRef]
66. Della Porta, M.G.; Travaglino, E.; Boveri, E.; Ponzoni, M.; Malcovati, L.; Papaemmanuil, E.; Rigolin, G.M.; Pascutto, C.; Croci, G.; Gianelli, U.; et al. Minimal morphological criteria for defining bone marrow dysplasia: A basis for clinical implementation of WHO classification of myelodysplastic syndromes. *Leukemia* **2014**, *29*, 66–75. [CrossRef]
67. Jing, Y. Research on the Application of Artificial Intelligence Natural Language Processing Technology in Japanese Teaching. *J. Phys. Conf. Ser.* **2020**, *1682*, 012081. [CrossRef]
68. Joshi, D.S.; Risodkar, Y.R. Deep Learning Based Gujarati Handwritten Character Recognition. In Proceedings of the 2018 International Conference on Advances in Communication and Computing Technology (ICACCT), Sangamner, India, 8–9 February 2018; pp. 563–566. [CrossRef]
69. Ma, S.; Chen, C.; Khalajzadeh, H.; Grundy, J. Latexify Math: Mathematical Formula Markup Revision to Assist Collaborative Editing in Math Q&A Sites. *Proc. ACM Human–Comput. Interact.* **2021**, *5*, 403. [CrossRef]
70. Ling, J.; Rush, A. Coarse-to-Fine Attention Models for Document Summarization. In Proceedings of the Workshop on New Frontiers in Summarization, Copenhagen, Denmark, 7 September 2017; pp. 33–42. [CrossRef]
71. Névéol, A.; Zweigenbaum, P. Expanding the Diversity of Texts and Applications: Findings from the Section on Clinical Natural Language Processing of the International Medical Informatics Association Yearbook. *Yearb. Med. Inform.* **2018**, *27*, 193–198. [CrossRef] [PubMed]