

A DEEP LEARNING-BASED APPROACH FOR URDU AND ENGLISH HANDWRITTEN TEXT RECOGNITION

Nazia Rabnawaz¹, Aiesha Ahmad², Humera Batool Gill³, Muhammad Ahsan Jamil^{*4},
Naeem Aslam⁵

^{1,5}NFC Institute of Engineering & Technology

²National College of Business Administration & Economics

³The Women University Multan

^{*4}MNS University of Agriculture Multan

^{*4}ahsan.jamil@mnsuam.edu.pk

DOI: <https://doi.org/10.5281/zenodo.21373933>

Keywords

Transformer-based models, multilingual, Optical character recognition (OCR), Urdu OCR, document analysis, and performance evaluation

Article History

Received: 01 January 2026

Accepted: 17 March 2026

Published: 31 March 2026

Copyright @Author

Corresponding Author: *

Muhammad Ahsan Jamil

Abstract

The difficulties presented by low-resource languages are significant when it comes to digitizing written material. Due to the simple linguistic resources that mostly abound within such languages, establishing efficient systems for accurate optical character recognition is an issue that calls for special consideration. Besides highlighting the importance and need to specialize on such languages, this paper introduces ViLanOCR, which is an innovative multilingual OCR system tailored specifically towards Urdu and English languages. The system takes advantage of sophisticated multilingual transformers within its language models that result in greater performances than other approaches, which pose considerable difficulties due to the nature of low resource languages. On the Urdu UHWR testing set, the proposed system achieves state-of-the-art performances with a CER measure of 1.1%. The testing findings show how effective the suggested method is, outperforming Cutting-edge beginnings in the digitization of Urdu handwriting.

INTRODUCTION

After the advent of digital technology, optical character recognition (OCR) has been thoroughly researched (Rao et al., 2016). OCR's goal is to convert scanned text images into a machine-readable format so that massive amounts of unstructured handwritten or printed text can be digitized and analyzed. In the 1960s and 1970s, researchers tried to recognize characters using manually created features and rule-based techniques, which led to the development of

optical character recognition (OCR). OCR systems have been greatly improved by recent developments in deep learning. The first generation of deep learning-based OCR models significantly increased recognition accuracy and flexibility across a range of scripts and writing styles by combining Long Short-Term Memory (LSTM) networks (Graves) with Convolutional Neural Networks (CNNs) (Chauhan, Ghanshala, and Joshi). This design was greatly enhanced by the creation of Region-Based Convolutional

Neural Networks (RCNNs) (Bao et al.), which employed region proposal techniques to improve feature extraction and object localization. The development of Region-Based Convolutional Neural Networks (RCNNs) (Bao et al.), which used region proposal techniques to better feature extraction and object localization, significantly improved this design. Despite these advancements, one significant drawback of these models was the need for a different language model for post-processing. The main reason for this is that they function at the character level instead of the word level. In addition to adding computing complexity, the addition of an external language model raises the possibility of language model-related errors spreading. Transformer-based models for OCR, namely Transformer-based OCR (TrOCR), were created as a result (Li et al., 2021). This substitutes a pre-trained Vision Transformer for the CNN layers and generates text using a pre-trained language model. Compared to earlier generations, this has demonstrated better accuracy and speed performance.

OCR still confronts a number of difficulties despite its achievements, especially when it comes to multilingual documents. Large multilingual datasets are challenging to handle due to the aforementioned drawbacks of earlier methods as well as their lack of scalability and parallelizability. Furthermore, the previous training of language system, that might not every time accurately represent the language and targeted document scripts determines the accuracy of TrOCR models. This issue is most evident with non-Latin scripts like Urdu and Arabic, where OCR conventional models exhibit subpar performance.

This research paper examines the digitalization of multilingual documents, particularly for low-resource languages. By examining current methods, instruments, and technologies employed in the process, it seeks to explore the opportunities and difficulties associated with digitizing low-resource languages. Next, we provide a way to solve the issue of digitizing handwritten multilingual medical prescriptions that include text written in Urdu and English, a language with limited resources. By putting out a model designed especially for dual-language text retrieval, we

overcome shortcomings of the earlier OCR methods. In addition to existing publicly accessible Urdu and English datasets, we employ machine generated script multilingual medical recipe data set record to illustrate the feasibility of our approach. Since the medical field necessitates the precise and effective digitalization of data found in handwritten prescriptions, it is an especially fascinating field to research.

The article is organized as follows for the remainder of it. A thorough survey of the literature on current OCR techniques is provided under "Related Work," with an emphasis on bilingual and multilingual approaches. The architecture and design of the suggested ViLanOCR model are explained in Section 3: Proposed Approach. Section 4: Experiments explains the datasets, experimental setup, and evaluation technique. Results and Discussion the experimental findings are presented and analyzed within Section 5, along with comparisons with current methods (DOI: 10.7717/peerj-cs, PeerJ Computer Science, 1964 Cheema et al., 2024). Additional trials carried out to confirm ViLanOCR's performance are covered in Section 6: Extended Study. The main findings of this study are summed up in Section 7: Conclusion and Future Directions, which also suggests possible directions for future research.

RELATED WORK

There are a number of well-established techniques for digitizing English-language documents, many of which can be used as a basis for creating techniques for digitizing multilingual materials. However, there are many difficulties in digitizing bilingual writings that include languages with different scripts, like Urdu and English. The basic distinctions between the writing systems, which require the identification and processing of various character sets and structural patterns, are a major challenge. Furthermore, model training is made more difficult by annotated data languages with low-resource or Lack of Parallelism, which restricts the creation of reliable OCR systems that can correctly handle such multilingual documents. Google created and maintains the popular open-source OCR system Tesseract (Rice, Jenkins, &

Nartker). Through a series of steps which involves optical character, layout analysis, character segmentation, text line extraction, and image preprocessing, it uses a variety of sophisticated algorithms and techniques to conduct text recognition. With the introduction of version 4, Tesseract experienced, a dramatic change, incorporating the latest LSTM engine with enhanced efficiency. Text recognition tasks are a good fit for LSTM, a kind recurrent neural network (RNN) which is excellent at processing sequential input. Because of LSTM engine's aggregation, Tesseract was able to manage complicated recognition mistakes, layouts of text more skillfully, and get more accurate results than their predecessors (Smith, 2016). The authors also attempted to modify Tesseract OCR, which is currently driven by LSTM, for use scenarios involving multiple languages. However, other than Latin languages, had poor Character Error Rates (CER) and Word Error Rates (WER) character error rates (CER) (with the LSTM + Dictionary approach, e.g., 25% WER for Hindi.), especially Indic language.

The EasyOCR tool is incredibly flexible for multilingual OCR workloads because it supports more than 80 languages (JaidedAI, 2023). This integrates a number of preprocessing techniques, including grayscale transformation along with other pertinent procedures, to improve the quality of input photos prior to text extraction. Furthermore, the character region awareness for text detection (CRAFT) method is applied successfully Baek et al. (2019). Convolutional recurrent neural networks (CRNNs) are used in EasyOCR'S recognition model. The CRNN labels sequences using Connectionist Temporal Classification (CTC) and long-short term memory. Even though long-short term memory (LSTM), which is a type of RNN, takes care of the sequencing aspect, Connectionist Temporal Classification (CTC) is responsible for tagging the Unlabeled sequencing datasets. Together, this makes it possible for EasyOCR to be capable of text extraction and recognition within a visual image. Despite developers' assertions of incorporates Urdu, EasyOCR fails to produce good result for the methods that we use.

According to Du et al. (2020), PaddlePaddle OCR (PPOCR), an open-source OCR solution that has proven to be a strong contender with its superior performance capability, versatility, and broad spectrum of supported languages. It uses state-of-the-art deep neural networks like Faster R-CNN, CRNN, and Mask R-CNN to accomplish tasks of text detection, recognition layout analysis, and recognition respectively. The flexible architecture, adaptable setup alternatives, as well as seamless integration of PPOCR have rendered it favored by academics and developers. Due to its flexible configuration options and smooth integration of PPOCR with modular framework has made this solution highly preferred by both researchers and developers. PPOCR holds a notable promise in the case of multilingual OCR tasks and supports a vast range of different scripts and alphabets like Arabic, Chinese, and others like Latin and Cyrillic alphabets either.

Nevertheless, its framework and training mechanisms enable language adaptation and therefore are suited for a bilingual OCR task. Li et al. (2021) provide TrOCR, a novel OCR architecture that makes use of pre-trained transformers. With this method, the authors do away with CNNs, which are typically used for OCR jobs involving picture feature extraction. Whereas, they use pre-trained vision transformers based

on self-attention techniques commonly used in tasks that involve natural language processing (NLP). Incompatible CNNs, which operate on the whole input image matrix, vision transformers divide the image into patches. These patches are treated as tokens and passed as input to the vision transformer model. The model creates embeddings that depict the input picture in the latent space in later layers by learning the weights of attention for each token. OCR-specific input-output pairs are used to jointly fine-tune the Vision Transformer and the language model during autoregressive training. CNN-LSTM architectures, which were widely used in earlier OCR systems, have limited scalability and were challenging to represent long-term dependencies since LSTMs are sequential. On the other hand, TrOCR uses a language model based on

RoBERTa (Liu et al.), which provides better contextual understanding and parallelization, resulting in better text recognition performance. TrOCR considerably streamlines the OCR pipeline by removing the requirement for a separate post-processing language correction module by integrating the language model directly into its architecture. The model establishes itself as one of the most sophisticated OCR systems now in use by achieving state-of-the-art performance on English handwritten text recognition tasks. However, compared to CNN-based methods, TrOCR's primary flaw is its computational overhead because it uses two massive transformer models one for vision and one for text generation which much lengthen inference times. Furthermore, the quality and linguistic diversity of the pre-training data utilized for the language model have a significant impact on its accuracy, making it difficult to adapt to non-Latin or low-resource scripts.

Urdu handwriting dataset is an important resource for the research of recognition. Text translation is performed by using an image-to-sequence framework method, which contain an improved Bidirectional Long Short-Term Memory network termed BiLSTM and Convolutional Neural Network termed CNN. To reduce loss of detail and preserve fine detail, the researchers replaced the standard max pooling operation in hidden layers with the addition of minor detail by concatenation, which is the greatest aspect of the CNN structure proposed by the researchers. Character Error Rate or CER was the measure by which system performance was evaluated, and on the UHWR dataset, they obtained staggering results of 7.35% CER. Although it provided promising results and there was scope for cross-lingual work, its application was only on text line level and not on document digitization. Moreover, since this requires high resolution pictures, combined with a special language system for error correction, its applicability in more general, resource-constraint situations can be questioned. Recently, a new method for the digitization of multilingual handwritten numbers was proposed by authors in the article, Fateh, Fateh, and Abolghasemi (2021). According to the authors,

the recognition of handwritten numbers has gained prominence, especially in relation to image processing, where the recognition of multilingual numbers has become more essential with each passing day.

Due to the variability of writing styles and handling languages of different scripts, it is hard to develop such structures. To solve these issues, the authors developed a language-independent methodology that basically has two parts, which are the Language Recognition (LR) and Digit Recognition (DR) components of this. The LR model classifies the input samples according to their language with a CNN-based methodology for language adaptation. After the identification of the language, the LR model forwards the sample to the respective DR the model. The DR model focuses more on digitizing the handwritten word. Its CNN structure is pretty similar to that of the LR model. Transfer learning methods to get better performance, along with methods to improve the quality of images, are applied. The authors have applied their solution on an extensive variety of six languages that comprise English, Kannada, Chinese, Persian, Arabic, and Urdu. Experimental results show state-of-the-art performance, with an average accuracy of as high as 99.8%. Research findings show The First, one of the key advantages of the method is that it is easily scalable, which provides the ability to add new languages without considerable rewriting of the whole system. As important as it is to note, the suggested method will be restricted to only numbers digitization and not to other kinds of handwritten text. Moreover, the LR model may fail to identify between languages which have identical number patterns. The written method of the number "3" precisely the same in each language like in Arabic, Persian, and Urdu.

Methodology

This section presents ViLanOCR, our innovative technique for handwritten bilingual text extraction from documents. By altering a multilingual vision language transformer model, it effectively extracts text from images in both Urdu and English. The design and operation of ViLanOCR are examined in this part, with a focus

on its primary elements and methods. The visual representation of the architecture is shown in Fig. 1.

We concentrate on the field of handwritten medical prescriptions in order to verify the effectiveness of ViLanOCR. Diagnoses, patient names, ages, medication names, and dosages are just a few of the many details included in these documents, which are frequently provided in both Urdu and English. We utilize the CER measure to assess the accuracy and robustness of our method. The percentage of wrong characters (insertions, deletions, and substitutions) in the OCR output relative to the ground truth is used to determine the CER, a metric for text recognition accuracy. The model transcribes handwritten text more accurately when the CER is lower. The development of a multilingual-based OCR system is a challenging task, especially for languages such as Urdu that have few resources. The use of huge-scale transformer-based designs, which have demonstrated remarkable performance on well-resourced languages, is limited by the lack of publicly accessible handwritten Urdu data.

To overcome the issue of a lack of available data, we combined self-submitted data from 30 individuals. We also included publicly available datasets like the printed Urdu vocabulary dataset in addition with self-submitted data. We also included NUST-UHWR (Ul Sehr Zia et al., 2022) and PUCIT-OHUL (Anjum & Khan, 2020, 2023),

as well. This combination of different datasets allows the model to be robust as well as adaptable. However, to make our dataset diverse in order to ensure high accuracy results are achieved rather than just relying on the number of available datasets, we make use of several augmentation techniques like pixelation, Gaussian noise, quantization noise, salt noise, pepper noise, as well as shear.

Our method requires the combination of two effective models: Swin Encoder developed by Liu et al. (2021), and mBART-50 Decoder developed by Tang et al. (2020). For this method, Swin Encoder will be primarily used for the feature extraction part. It is skilled at capturing minute details and thus ensures that the features are effectively extracted. One of the most interesting characteristics of the Swin encoder is the sliding window mechanism. Thanks to this mechanism, the model can analyze the input image in a sequential manner.

Swin transformers are very useful in dense tasks such as object detection and segmentation because, in comparison to vanilla vision transformers, they make use of the shift window attention, thereby reducing the attention mechanism to operate within a small proximity. Due to the fact that even small changes can have a significant impact on handwritten language, our problem is likewise dense and demands pixel-level comprehension.

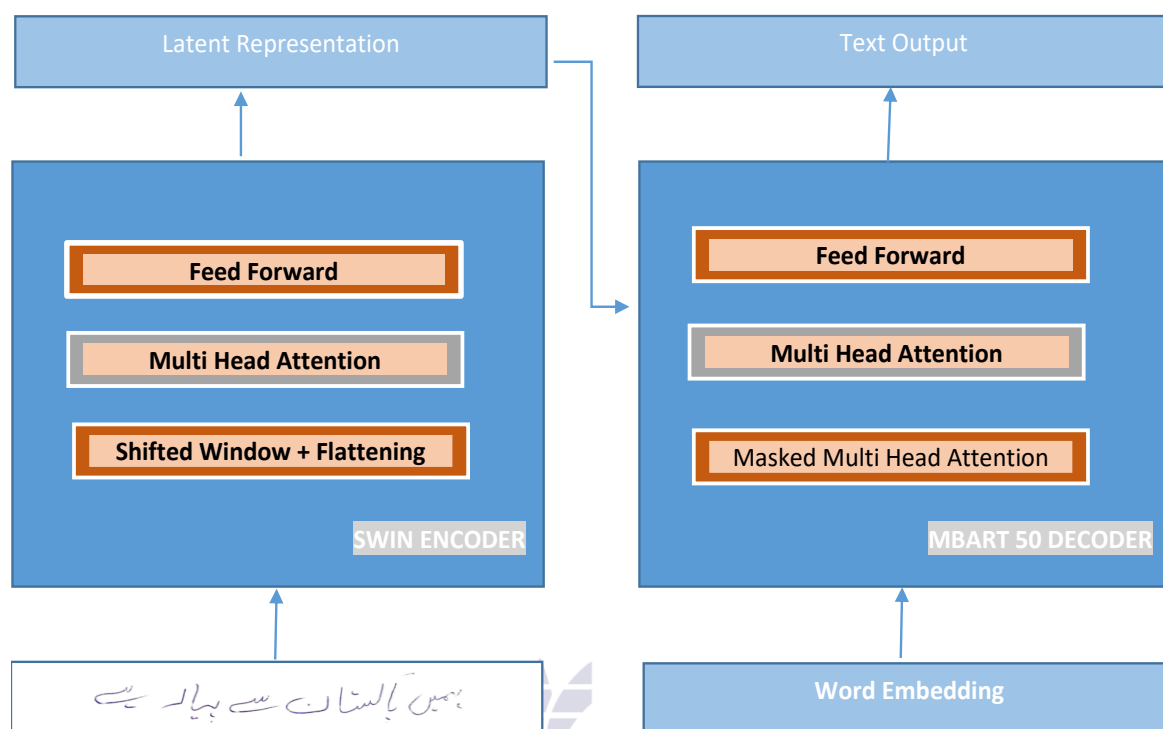


Figure 1 ViLanOCR architecture



Swin transformers are excellent for dense tasks like object detection and segmentation because, in contrast to vanilla vision transformers, they employ shift window attention, which reduces the attention mechanism to work in a small proximity. Due to the fact that even small changes can have a significant impact on handwritten language, our problem is likewise dense and demands pixel-level comprehension. By localizing attention into smaller regions and capturing slight changes in text, the Swin Transformer's shifting window partitioning successfully addresses the subtle variances in handwriting. Each non-overlapping patch in the input image is regarded as a token. For instance, a patch of size 4×4 provides a feature dimension of 48 ($4 \times 4 \times 3$). The features of these patches can be directly garnered by their pixel value representation. These features require projection in the latent space. They require a

linear embedding as the way for this projection in the latent space. These tokens need to be transformed by the Swin Transformer blocks, which can be considered variants of the usual Transformer blocks. In order to reduce the number of tokens and double the feature dimension, each patch merging layer combines the features of nearby patches and applies a linear layer to these combined features. To halve the number of tokens and increase the features by two times, each patch merging layer computes the features of patches in the neighborhood and applies an additional linear transformation on them. Like the layers in regular CNN models such as either VGG nets or ResNets, this process generates a hierarchical representation of the given image. As discussed earlier, Swin Transformers differ in that they use shifted window partitioning. In simpler terms, this implies that the Transformer

performs regular window partitioning in one block and somehow shifts the windows in the next block. By creating links between nearby windows, this shifting improves the model's capacity to capture linkages between various areas of the image. In tasks that require an in-depth, contextual understanding of the visual, such as handwritten English in our case, this process is thoroughly necessary.

In addition, the self-attention computation in the Swin Transformers includes relative location bias. It increase the ability of handling complex image information, by allowing an understanding of the relative positions many patches in the window occupy particularly challenging spatial reasoning like the case of the Urdu handwritten letter.

Another reason for using Swin encoder is that it is more efficient than the traditional Multi-Head Self-Attention (MSA) approach. The computer complexity of the traditional global MSA is quadratic with respect to the number of patches that are found in the image. Swin Transformer adopts the-window based MSA. Here, the self-attention value for every patch is computed within a local window of $M \times M$ size. The complexity of Swin Transformer is linear. Subsequently, Swin Transformer is much more scalable and efficient for dense tasks like high-resolution images or high-resolution images handwriting.

On the other hand, the mBART-50 decoder is very proficient at dealing with a diverse array of scripts and languages. The application of multi-head attention, a method which allows a model to pay attention to a number of regions of a source image simultaneously, is a characteristic application of multi-task learning and is a distinctive feature of mBART-50 decoders. This becomes particularly important within multilingual text extraction, as that would allow it to work simultaneously with the Urdu and English parts of a document, irrespective of whether these parts are spatially separated or mixed. The distinctive ability to function as a denoiser is very important for improving the accuracy and reliability of our ViLanOCR system. As a function of its role as a denoiser and a type of AutoEncoder, mBART-50 has been preconditioned. A collection of monolingual documents and the texts of a

combination of N-language datasets are included in this pre-conditioning process. Two types of application of noise: random span and permutation of order are diverse and are incorporated at random points within the preconditioning phase. Both of these plays a pivotal part within giving a robust and complex understanding of language quirks—a very important commodity in decoding multiple scripts and languages—a prerequisite for ViLanOCR. Within this particular context, noise is deliberately added to text regions of a particular image through random masking. For experimenting with replicating noise or distortion often discovered in a diverse array of hand-written texts within the actual context of human utilization, this particular function of a mask is successful at camouflaging certain regions. The responsibility of mBART-50 decoders lies within teasing back or decoding these masked sections within order to determine the masked or deleted texts. "Model Experimentation" gives a detail-oriented description within how Vision Language-based OCR (ViLanOCR) operates.

EXPERIMENTS

The experimental findings from ViLanOCR are elaborately presented, in this section. The efficiency of the proposed model will be carefully evaluated using different datasets. These include the self-created synthetic medical prescriptions dataset created by the authors themselves. Others include publicly available datasets.

These carefully selected and combined datasets offer a thorough evaluation of our suggested models in practical situations. Most especially, adding our own Medical Prescriptions Dataset further diversifies the scope of our evaluation and epitomizes our power in adequately handling bilingual situations. Our model's performance on multilingual Data entry inaccuracies is overcome by using our own dataset that increases the adaptiveness of our method and allows for serious evaluation. We focus on the following aspects explain in detail our training and assessment procedures.

English datasets

To create such an extensive corpus for solving the task of English text recognition, we merged publicly available handwritten and printed texts. These included HW_wiki, which contains over two million handwritten words obtained from different sources such as academic papers, Wikipedia, among others. The paper by Davis et al. (2020) describes the HW_wiki corpus. We also employed the IAM dataset (Marti & Bunke, 2002), which includes handwritten text in English samples are taken from a number of writers and styles, including letters, diaries, and forms, to further increase the diversity of the English dataset. We also used the SROIE dataset (Huang et al., 2019), which includes a variety of printed and handwritten text samples of invoices, receipts, and different business documents. This made it possible for us to further develop our models on greater value and authentic text recognition data. We used the TRDG synthetic dataset (Belval, 2019), which includes more than two million samples of synthetic printed content.

Urdu datasets

Similarly, for the Urdu dataset, we specifically selected both printed and handwritten publicly available datasets. For handwritten text, the NUST-UHWR (ul Sehr Zia et al., 2022) and PUCIT-OHUL (Anjum & Khan, 2020, 2023) datasets were utilized; for printed text, the UPTI (containing 0.5 million samples) and UPTI v2 (containing 1 million samples) datasets were utilized. In addition to making sure the data was properly labeled and prepared for our research, we also took care to choose datasets that are representative of the Urdu language and varied.

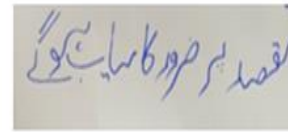
Augmentation

Our dataset size was expanded from 20 to 200 GB by applying different techniques of image augmentation to further enhance and expand dataset.

Our techniques for image augmentation aim to improve and alter the input photos' visual features. They utilize a range of techniques to produce a variety of outcomes. To create blur effects, we implemented filters like general blur, box blur, and Gaussian blur. By using smoothing methods these filters create the images appear softer. By adjusting the blur radius, several levels of blurriness can be achieved, offering flexibility in the desired outcome. We used contrast adjustment to alter the images' contrast. Lastly, we employed grid augmentation to give the pictures grid-like patterns. The horizontal grid function adds horizontal lines to the photos, and the vertical grid function adds vertical lines. Because the line widths and spacings are chosen at random, the grid's look is flexible. Figure 2 displays every image augmentation that was utilized.

Apart from providing an approach with a model that can recognize text in both languages, we have also developed models specifically designed to recognize text only in Urdu, in the form of one line and multiple lines. Moreover, we have optimized the TrOCR, one of the best OCR models available in the market, using our extensive dataset. Finally, depending on the context of our solution, we have also implemented text extraction components for Chinese and Japanese languages. We evaluated several methods and techniques to identify a solution to cater to our requirement of extracting both single-line and multi-line text comprising of both Urdu and English languages. In order to ensure that the student learns from the instructor through attentiveness, we first tried using the Data-Efficient Image Transformer (DeiT) (Touvron et al., 2021), which uses a teacher-student technique to train the model from a smaller collection of labeled data. This basically attempts to reduce the size of the model and eliminate the pre-training component of transformers, which are pre-trained utilizing costly infrastructure and hundreds of millions of photos. Unfortunately, in our use case, we discovered that it did not converge.

(a)



(b)



(c)

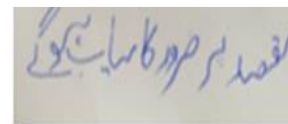


Table 1 Number of parameters for each model.

Model	# of parameter
TrOCR	558M
Pytesseract	-
EasyOCR	8.3M
PPORC	8.1M
CRNN	4M
BiLSTM	-
CNN Transformer	-
ViLanOCR English	497M
ViLanOCR Urdu	497M
ViLanOCR-Multi	497M

RESULTS AND DISCUSSION

The study's objectives were to use different deep learning models for text localization in order to

digitize bilingual handwritten text (Urdu + English) on documents and evaluate the effectiveness of our new OCR system, ViLanOCR,

against a number of established OCR models. The results are shown and discussed below.

Table 2 shows the CER as a percentage for various OCR models on many benchmark datasets, including our new ViLanOCR. A variety of datasets were used for the evaluation, including IAM, HWwiki, TRDG-SYN, SROIE, NUST-UHWR, PUCIT-OHUL, UPTI, and our own artificially created dataset, Medical-SYN, which consists of handwritten medical prescriptions in both Urdu and English.

Our recommended ViLanOCR method shown excellent results, especially on datasets including Urdu text. It yielded CER values of 1.1%, 5.22%, 6.3%, and 6.77% for the NUST-UHWR, UPTI, PUCIT-OHUL, and Medical-SYN datasets, consecutively. These findings illustrate ViLanOCR's advanced capability to accurately recognize and extract Urdu characters from diverse document sources. Additionally,

ViLanOCR-Multi demonstrated competitive performance on English text extraction tasks, with CER values of 6.12% on the HW wiki dataset and 5.92% on the IAM wiki dataset. The system's ability to efficiently handle both English and Urdu content is demonstrated by this. With CER values of 3.42% and 2.86%, respectively, TrOCR showed notable performance on the IAM and HW wiki datasets among the benchmark OCR models. With a CER of 56.39% on Medical-SYN and 78.23% on NUST-UHWR, it did incredibly poorly on Urdu text extraction, though. Although the performance of the previously described benchmark OCR models varied, our ViLanOCR system stands out as a reliable and efficient method for extracting text in both Urdu and English. Its capacity to outperform well-known OCR models on datasets with Urdu content highlights its important addition to the field, as seen in Fig. 3.

Table 2 Character error rate (CER), in percentage, on various benchmark datasets, and our own synthetically generated medical prescriptions datasets, Medical-SYN.

	English				Urdu			Bilingual
Model	IAM	HW wiki	TRDG-SYN	SROIE	NUST-UHWR	PUCIT-OHUL	UPTI	Custom-Medical
TrOCR	3.42	2.86	1.38	2.17	78.23	-	-	56.39
Pytesseract	69.2	-	-	-	-	-	-	74.02
EasyOCR	56.7	-	-	-	97.89	98.17	-	60.6
PPORC	32.32	29.3	5.56	3.11	96.25	-	-	45.31
CRNN	-	-	-	-	95.72	98.45	94.19	92.13
BiLSTM	-	-	-	-	94.19	97.01	95.21	90.98
CNN Transformer	48.14	-	-	-	59.28	62.63	60.53	76.20
ViLanOCR English	3.79	3.64	1.24	1.92	-	-	-	13.82
ViLanOCR Urdu	-	-	-	-	1.1	5.22	6.3	21.23
ViLanOCR-Multi	5.92	6.12	3.26	2.31	3.21	9.85	9.92	6.77

The suggested ViLanOCR system outperforms a number of well-known OCR models in a number of benchmark datasets, especially its Multilingual variant, ViLanOCR Multi. It is crucial to recognize that the efficacy of each OCR model

may differ based on the dataset's properties and the type of text that needs to be extracted.

EXTENDED STUDY

We suggest an end-to-end method called MEDIGIZE, which is a logical progression of the

ViLanOCRframework, in order to improve upon ViLanOCR while keeping our medical case

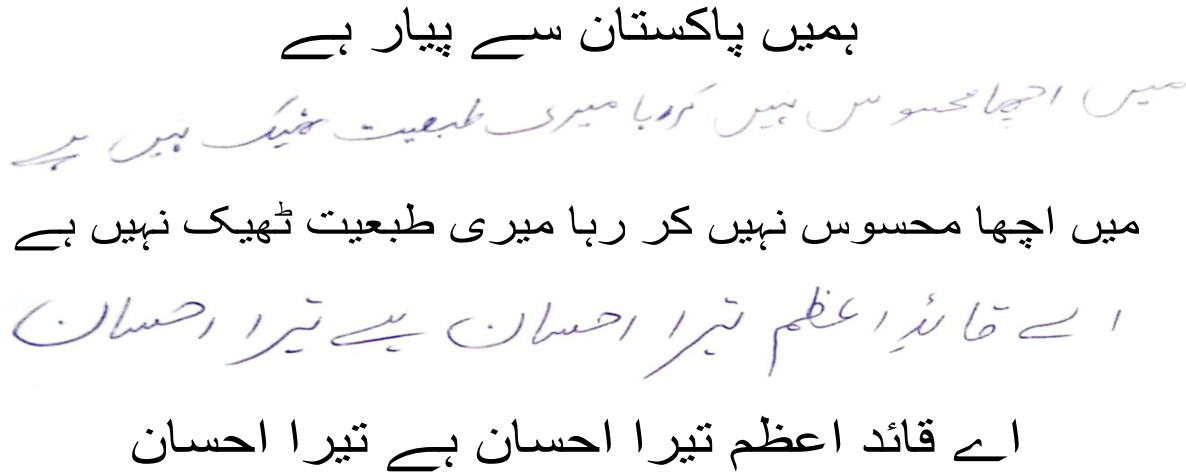


Figure 3 Inference results on images not part of the training set.

study in mind. MEDIGIZE stands out as a convincing solution since the problems of our initial case study, medical prescription digitization, call for an all-encompassing strategy. MEDIGIZE covers the entire digitization pipeline, including text localization, text extraction, and entity extraction, by incorporating ViLanOCR's capabilities. This comprehensive strategy is based on the knowledge that there are unmatched benefits to converting handwritten documents, especially the medical prescriptions depicted in Fig. 4, into organized and easily accessible digital formats.

Each of the three crucial phases that make up the MEDIGIZE framework makes a substantial contribution to the process as a whole. Handwritten sections of interest are found during the first text localization stage, opening the door for effective further processing. After that, these localized areas move on to the text extraction stage, where ViLanOCR is used to provide precise and adaptable text extraction from a variety of scripts and languages. Despite being unstructured in its unprocessed state, this retrieved text has enormous potential for further study. The final step is entity linking, where a sophisticated NER

model is applied to the digital text. The retrieved text is given structure by this categorization procedure, which separates its different elements, including patient information, diagnoses, prescription drugs, dosages, and more. Fig. 5 depicts this architecture.

The discovery that an end-to-end pipeline bridges the gap between valuable information and simple text extraction was the primary driving force behind its creation. ViLanOCR is a skilled text extraction tool, but switching to a structured format using NER models opens up many advantages. In addition to making data management and accessibility easier, structured data facilitates intelligent analysis and decision-making. MEDIGIZE improves data-driven procedures, increases the effectiveness of data storage, and eventually helps to improve medical document management by rescuing relevant information from the textual debris.

YOLOv5, YOLOv6, YOLOv7, and YOLOv8 are some of the refined YouOnlyLookOnce (YOLO) (Redmon et al., 2016) models that we used in our study on handwritten text identification. Our bespoke medical prescription dataset, which includes more than 50,000 samples (after

augmentation) of both Urdu and English handwriting, was used to thoroughly evaluate these models. After thorough testing and analysis, we observed that YOLOv5 achieved the highest average mean precision (mAP) of

0.897 among all evaluated versions. We additionally employed various data augmentation techniques, including scaling, rotation, and noise addition, to enhance the accuracy of our system's handwritten text detecting.

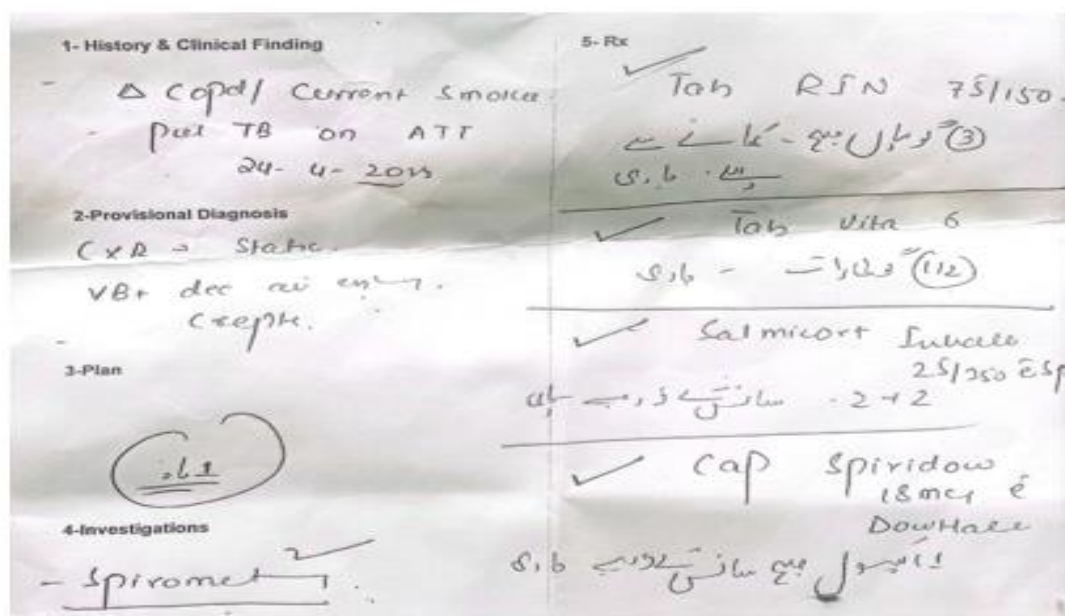


Figure 4 Medical Prescriptions dataset. | Photo: Author's own.

We observed that these strategies, which encompassed a comprehensive evaluation of various YOLO models alongside the implementation of methods to augment data and transfer learning to improve precision as well as effectiveness, contributed to the enhancement of the variety and complexity in handwritten text recognition.

We decided to adopt YOLO methodology for object detection represents because this is the most sophisticated methodology in trend, therefore we have chosen to implement it in our model for the

localization of text. YOLO is a current object identification system renowned for its exceptional accuracy and rapid processing speed, rendering it widely utilized in various computer vision applications. The fact that YOLO is an end-to-end model that accurately forecasts the bounding boxes and class probabilities for every item in a single forward neural network run is one of its main features. This indicates that YOLO's real-time processing speed makes it appropriate for use cases like ours that need for quick and precise object recognition.

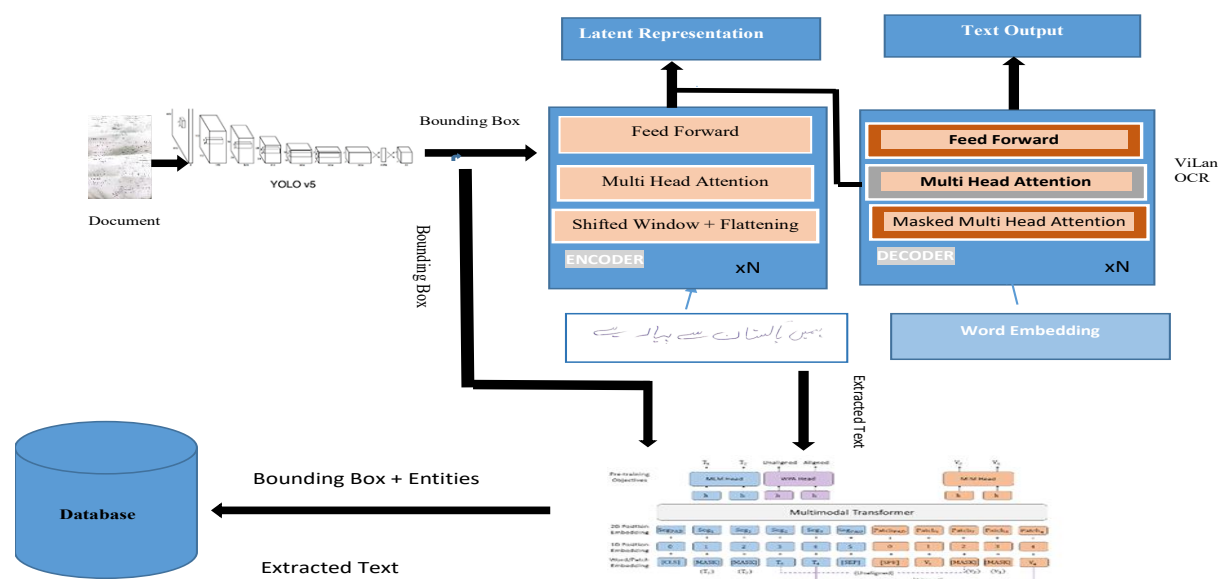


Figure 5 Medigize a pipeline to digitize the extraction of entities from bilingual medical prescription.

ViLanOCR, our previously stated OCR, is utilized as a modular plug-in OCR for text extraction. Because Medigize's entire workflow is modular, experimenting with various tactics is easy.

Extraction of Entity

It is crucial to comprehend the context and relationships between various entities in order to extract useful information from the recovered text. The layout-based language model is essential for determining the relationships between retrieved text items and the document's structure.

We are using a potent pre-trained language model called LayoutLMv3 (Huang et al., 2022), which was created especially for document artificial

intelligence applications. It can accurately predict the class of the extracted text because it has been trained on a large dataset of text and layout data. By adding and optimizing LayoutLMv3 into the pipeline we have created, we can obtain data such as diagnosis, person name, age, drug name, dosage, and more, and identify the relationships between them. The resulting organized data is easily examined and understood by medical professionals. Furthermore, the model's OCR foundation makes it perfect for our challenge because we already use our own cutting-edge OCR for text extraction.

Model	Precision	Recall	F1
YOLO v3	0.857	0.811	0.833
YOLO v4	0.8515	0.8045	0.827
YOLO v5	0.897	0.874	0.885
YOLO v7	0.836	0.777	0.805

Table 3 Text detector results for identifying handwritten text on document.

By employing this procedure, we can also providing the user with information on the locations of the entities that have been selected from the text and also we are able to accurately extract important information from medical records. This information can help medical staff identify potential misclassifications of entities when visualized for each prescription.

Results

The performance metrics, namely Recall and Precision, reported by different text analyzers to identify handwritten text from documents are presented in Table 3. With an outstanding Precision value of 0.897 and Recall value of 0.874, YOLO v5 established itself as having the best possible performance between models tested. This suggests that YOLO v5 is much better equipped to make correct annotations with regard to relevant written parts in papers, having superior recall value. At the other end, YOLO v7 was better compared to YOLO v3 and YOLO v4. YOLO v7 illustrated that it was far inferior to other models tested related to text localization, having low Precision and Recall values amongst models tested. To make accurate predictions, we used

LayoutLMv3, developed by Huang et al. (2022) above, that is multi-modal, capable of grasping both image properties from documents and information available from our above-mentioned OCR. The multi-modal approach named "LayoutLM" demonstrated its efficiency in handling complex layouts from documents, being potentially viable in NER, which exploited both text and image information to make correct entity predictions.

Multimodal style used by LayoutLM gave the results that were utilized to show its effectiveness in dealing with complex document layouts and emerged as a promising approach for NER task because it addressed visual and text modal altogether for entity prediction. F1 score of 0.86 indicates that we were able to detect things in papers with a quite precise level of recall and precision because LayoutLM had the ability to handle both visual and text modal quite effectively.

Perhaps, despite the encouraging results, there could be room for further enhancement. Through the fine-tuning of the model using domain-level data, the ability to identify entities for varied contexts could be improved.

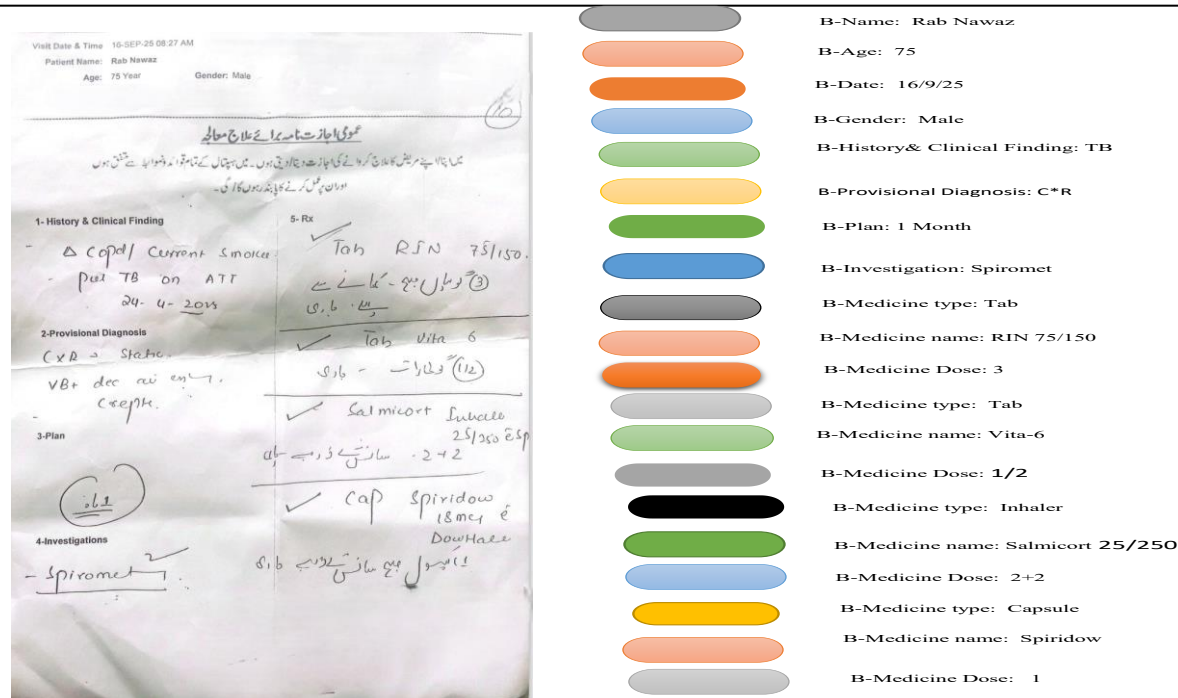


Figure 6 MEDIGIZE-based inference of a synthetic bilingual Handwritten medical prescription

According to Kim et al. (2021) and Davis et al. (2022), MEDIGIZE's modular three-stage methodology has a number of clear advantages over existing end-to-end methods.

I. Error Robustness and Mitigation: MEDIGIZE diminishes error propagation by segmenting the digitization process into many steps. End-to-end methods, on the other hand, are less dependable due to their susceptibility to error accumulation. Errors that happen during the text extraction stage could have a big impact on processing that comes after. Targeted mistake discovery, isolation, and correction are made possible by MEDIGIZE's modularity, which improves overall accuracy and dependability.

II. Optimization and Specialization: Each stage of a MEDIGIZE has to be differently configured in order to meet their different demands. Each stage of a MEDIGIZE model can be specifically optimized based on its different demands. End-to-end models, on the other hand, which may compromise optimization in the process because they need to perform well for multiple tasks. Optimization is enhanced by specialized optimization, which boosts the total efficacy of the

framework. The efficacy of each step is enhanced by the use of specialized optimization.

III. Lower processing costs: MEDIGIZE has been able to cut processing costs due to its modular design capabilities. The end-to-end approach increases processing complexity as all tasks need to be processed and handled by one model. MEDIGIZE reduces computational demands and enhances cost-effectiveness by distributing the burden among specialist models.

CONCLUSION AND DIRECTIONS FOR THE FUTURE

This study introduces ViLanOCR, a unique bilingual OCR that makes use of multilingual transformer-based language models. ViLanOCR makes it easier to maintain production-level models by precisely extracting text in both Urdu and English. In the past, each language needed its own OCR model, which meant that each model had to be managed and maintained separately.

However, management and maintenance at the production level are made easier by using a single OCR for both languages. Additionally, ViLanOCR establishes itself as the standard

model for Urdu OCR moving ahead by achieving state-of-the-art CER on a number of benchmark Urdu datasets. Additionally, it attains cutting-edge CER on bilingual (English + Urdu) OCR workloads.

Nevertheless, a few shortcomings in our findings present chances for future research to be even better. The employment of transformer-based models for OCR presents a number of difficulties, including the issue of language models experiencing hallucinations and high computing costs. Future research should examine methods for reducing transformer models without sacrificing performance and look at more reliable loss objectives to reduce hallucination errors in order to overcome these problems. Quantization or distillation of these big transformer models is one potential remedy.

Furthermore, recent developments in document AI have demonstrated potential methods for extracting key-value entities from document photos without the need for OCR. To effectively address the low-resource issue of digitizing handwritten medical prescriptions, these approaches are still in their infancy and will need more investigation and testing. The propagation of mistakes between modules is one of the issues in our pipeline that may be resolved if such techniques can be implemented. Furthermore, investigating OCR-free methods for information extraction from document photos could greatly improve digitization, especially when it comes to managing handwritten medical prescriptions and reducing mistake propagation across the pipeline.

REFERENCES:

- [1] Anjum, T., and N. Khan. "An Attention-Based Method for Offline Handwritten Urdu Text Recognition." 2020 17th International Conference on Frontiers in Handwriting Recognition (ICFHR), IEEE, 2020, pp. 169-174.
- [2] Anjum, T., and N. Khan. "CALText: Contextual Attention Localization for Offline Handwritten Text." Neural Processing Letters, vol. 55, no. 6, 2023, pp. 7227-7257. DOI: 10.1007/s11063-023-11258-5.
- [3] Baek, Y., B. Lee, D. Han, S. Yun, and H. Lee. "Character Region Awareness for Text Detection." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE, 2019, pp. 9365-9374.
- [4] Bao, Y., H. Li, X. Fan, R. Liu, and Q. Jia. "Region-Based CNN for Logo Detection." Proceedings of the International Conference on Internet Multimedia Computing and Service, ACM, 2016, pp. 319-322.
- [5] Belval, E. TextRecognitionDataGenerator. GitHub, 2019. Accessed 1 May 2023.
- [6] Chauhan, R., K. K. Ghanshala, and R. Joshi. "Convolutional Neural Network (CNN) for Image Detection and Recognition." 2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC), IEEE, 2018, pp. 278-282.
- [7] Davis, B., B. Morse, B. Price, C. Tensmeyer, C. Wigington, and V. Morariu. "End-to-End Document Recognition and Understanding with Dessurt." European Conference on Computer Vision, Springer, 2022, pp. 280-296.
- [8] Davis, B., C. Tensmeyer, B. Price, C. Wigington, B. Morse, and R. Jain. "Text and Style Conditioned GAN for Generation of Offline Handwriting Lines." arXiv, 2020. DOI: 10.48550/arXiv.2009.00678.
- [9] Dosovitskiy, A., et al. "An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale." arXiv, 2020. DOI: 10.48550/arXiv.2010.11929.
- [10] Du, Y., et al. "PP-OCR: A Practical Ultra-Lightweight OCR System." arXiv, 2020. DOI: 10.48550/arXiv.2009.09941.
- [11] Fateh, A., M. Fateh, and V. Abolghasemi. "Multilingual Handwritten Numeral Recognition Using a Robust Deep Network Joint with Transfer Learning." Information Sciences, vol. 581, no. 3, 2021, pp. 479-494. DOI: 10.1016/j.ins.2021.09.051.

- [12] Graves, A. "Long Short-Term Memory." Supervised Sequence Labelling with Recurrent Neural Networks, Springer, 2012, pp. 37-45.
- [13] Huang, Z., et al. "ICDAR 2019 Competition on Scanned Receipt OCR and Information Extraction." 2019 International Conference on Document Analysis and Recognition (ICDAR), IEEE, 2019.
- [14] Huang, Y., T. Lv, L. Cui, Y. Lu, and F. Wei. "LayoutLMv3: Pre-Training for Document AI with Unified Text and Image Masking." Proceedings of the 30th ACM International Conference on Multimedia, ACM, 2022, pp. 4083-4091.
- [15] JaidevAI. EasyOCR. GitHub, 2023. Accessed 2023.
- [16] Kim, G., et al. "Donut: Document Understanding Transformer without OCR." arXiv, 2021. DOI: 10.48550/arXiv.2111.15664.
- [17] Kumar, R., and R. Pal. "India Achieves WHO Recommended Doctor Population Ratio: A Call for Paradigm Shift in Public Health Discourse!" Journal of Family Medicine and Primary Care, vol. 7, no. 5, 2018, p. 841. DOI: 10.4103/jfmprc.jfmprc_218_18.
- [18] Li, M., et al. "TrOCR: Transformer-Based Optical Character Recognition with Pre-Trained Models." arXiv, 2021. DOI: 10.48550/arXiv.2109.10282.
- [19] Liu, Z., et al. "Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows." Proceedings of the IEEE/CVF International Conference on Computer Vision, IEEE, 2021, pp. 10012-10022. DOI: 10.1109/ICCV48922.2021.00986.
- [20] Liu, Y., et al. "RoBERTa: A Robustly Optimized BERT Pretraining Approach." arXiv, 2019. DOI: 10.48550/arXiv.1907.11692.
- [21] Marti, U.-V., and H. Bunke. "The IAM-Database: An English Sentence Database for Offline Handwriting Recognition." International Journal on Document Analysis and Recognition, vol. 5, no. 1, 2002, pp. 39-46. DOI: 10.1007/s100320200071.
- [22] Rao, N. V., A. Sastry, A. Chakravarthy, and P. Kalyanchakravarthi. "Optical Character Recognition Technique Algorithms." Journal of Theoretical & Applied Information Technology, vol. 83, no. 2, 2016, pp. 275-282.
- [23] Redmon, J., S. Divvala, R. Girshick, and A. Farhadi. "You Only Look Once: Unified, Real-Time Object Detection." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2016, pp. 779-788.
- [24] Rice, S. V., F. Jenkins, and T. A. Nartker. The Fourth Annual Test of OCR Accuracy. University of Nevada, Information Science Research Institute, 1995.
- [25] Smith, R. "Tesseract Blends Old and New OCR Technology." GitHub, 2016.
- [26] Tabassum, S., et al. "An Online Cursive Handwritten Medical Words Recognition System..." Scientific Reports, vol. 12, 2022, p. 3601. DOI: 10.1038/s41598-022-07571-z.
- [27] Tang, Y., et al. "Multilingual Translation with Extensible Multilingual Pretraining and Finetuning." arXiv, 2020. DOI: 10.48550/arXiv.2008.00401.
- [28] Touvron, H., et al. "Training Data-Efficient Image Transformers & Distillation Through Attention." International Conference on Machine Learning, PMLR, 2021, pp. 10347-10357.
- [29] ul Sehr Zia, N., et al. "A Convolutional Recursive Deep Architecture for Unconstrained Urdu Handwriting Recognition." Neural Computing and Applications, vol. 34, no. 2, 2022, pp. 1-14. DOI: 10.1007/s00521-021-06498-