BHCP-DOC: A Real-World Bilingual Dataset for Handwritten Clinical Prescription Detection and Recognition

Fatima Bukhari[1][0000−0002−1108−0312], Mickael Coustaty[2][0000−0002−0123−439X], Muhammad Muzzamil Luqman[2][0000−0002−9658−0833], Souhail Bakkali[3][0000−0001−9383−3842], Michael G. Madden[4][0000−0002−4443−7285], and Nadeem Iqbal Kajla[4⋆][0000−0001−8722−7790]

[1] Institute of Computing, MNS University of Agriculture, Multan, and National University of Modern Languages (NUML), Pakistan
fatima.bukhari@numl.edu.pk
[2] L3i, La Rochelle University, La Rochelle, France
mickael.coustaty@univ-lr.fr, muhammad_muzzamil.luqman@univ-lr.fr
[3] Univ Rennes, CNRS, IRISA - UMR 6074, Rennes, France
souhail.bakkali@irisa.fr
[4] School of Computer Science, University of Galway, Ireland
nadeem.kajla@universityofgalway.ie, michael.madden@universityofgalway.ie

Abstract. Although OCR has improved significantly for printed and Roman-script text, handwritten text, especially in doctors' handwriting, remains a challenge. This is because handwriting in prescriptions is often highly illegible, with letters and strokes frequently merging, word boundaries becoming unclear, and Urdu script adding further complexity due to its rich and interconnected character shapes. Progress has also been slow because large, high-quality annotated datasets for training deep learning models are limited. We curated BHCP-DOC, a real-world bilingual Urdu-English handwritten clinical prescription dataset containing 893 prescription images and 11,008 manually annotated text-line instances. The dataset captures challenging real-world conditions, including blur, low illumination, curved writing lines, complex backgrounds, illegible handwriting, stylized writing, occlusions, and mixed-script content. Each text line is annotated with a rectangular bounding box, transcription, and semantic category, such as medicine name, dosage, diagnosis, symptoms, or general text. We also report baseline results for text detection and text recognition on this dataset to support a fair comparison in the future. BHCP-DOC offers a clear benchmark and aims to speed up research on robust OCR for bilingual datasets and other low-resource handwritten scripts. The proposed dataset is publicly available at https://l3i-share.univ-lr.fr/2025BHCP-DOC/.

Keywords: Handwritten prescriptions · Bilingual handwritten text · OCR · Handwritten text recognition · Handwritten text detection · Handwritten clinical document

⋆ Corresponding authors: Nadeem Iqbal Kajla and Mickael Coustaty. Email: nadeem.kajla@universityofgalway.ie, mickael.coustaty@univ-lr.fr

1 Introduction

Recent research on handwritten text detection and recognition has expanded rapidly due to its importance in document analysis and the inherent challenges of handwriting, including variations in stroke formation and legibility. Deep learning has enabled substantial progress in both detection and recognition for widely studied scripts such as Latin, Chinese, and Arabic. However, bilingual handwritten datasets remain limited, especially in the medical domain, where clinical documents and prescriptions often contain inconsistent writing styles, abbreviations, and illegible handwritten text. Deep learning-based approaches have been investigated in several contexts, including the detection and recognition of documents in diverse styles, as shown in Figure 1, where documents contain clear and illegible English and Urdu handwritten text styles written by different clinicians. Urdu, as a low-resource language, remains comparatively underrepresented in OCR research, largely due to data scarcity and the limited availability of mature, high-performing recognition tools. We introduce the BHCP-DOC dataset, designed to capture the real-world complexity of clinical handwritten documents. Its difficulty stems from two primary factors: (i) the noisy and highly variable nature of handwritten prescriptions, and (ii) the coexistence of two languages (English and Urdu) and scripts within the same samples. Urdu differs fundamentally from Latin-based scripts in how words are formed [13]. Urdu characters have context-dependent shapes. Sometimes two or more letters also join together into a single new shape, often written smaller underneath the main consonant [7]. These combinations determine consonant sounds. Vowels are also flexible in placement and can appear below, above, after, and before consonants [11] where they merge to form new vowel sounds. As in Figure 1, Handwritten text by different clinicians, with messy and clear writing patterns. These properties, combined with the low-resource status of many non-Latin languages, create significant challenges for robust OCR [7], [23]. Recognizing handwritten text

(a) Different Clinicians' Urdu Handwritten Text Styles

(b) Different Clinicians' English Handwritten Text Styles

Fig. 1: Different Clinicians' English and Urdu Handwritten Text Styles

accurately is difficult mainly because there is not enough high-quality labeled data, which reduces how well trained models learn and work on new and unseen samples.

1. There is a significant and acknowledged lack of publicly available, high-volume datasets containing real-world images of real clinical prescriptions.
2. Existing datasets used for handwriting recognition (e.g., IAM, RIMES, or synthetic medical text generators) often fail to capture the unclear and missing parts of a doctor's handwriting, especially Urdu text.
3. Prescriptions involve technical medical terms and highly abbreviated instructions, which standard general handwriting recognition models are not trained to interpret.
4. The models exhibit low confidence scores for predicted words and a high CER and WER when processing authentic prescription images, effectively rendering the output unreliable for clinical use.

To address these limitations, this paper makes three main contributions:

1. We introduce BHCP-DOC, a real-world bilingual Urdu-English handwritten clinical prescription dataset containing 893 anonymised prescription images and 11,008 manually annotated text-line instances.
2. We provide rectangular bounding boxes, line-level transcriptions, and semantic labels for prescription text, including medicine names, dosage, diagnosis, symptoms, and general text.
3. We report baseline detection and recognition results using YOLO variants, CRNN-based models, attention-based recognition, and TrOCR.

Doctors' handwriting in clinical prescriptions is often difficult to recognize due to rapid writing, inconsistent letter forms, and domain-specific abbreviations [18]. Such ambiguity increases the risk of misreading medicine names or dosages, which can lead to medication errors and serious adverse outcomes, including death. OCR evaluation commonly relies on Levenshtein distance [21], which measures insertions, deletions, and substitutions between predictions and ground truth. For BHCP-DOC, visually similar words can differ in internal character order within character clusters, making direct string comparison unreliable without normalization. The core components are Urdu/English word segmentation and normalization. We apply the Urdu-English word segmentation to split predicted and ground truth texts, and use Urdu syllable reordering for normalization. We adapt only the syllable reordering necessary to standardize internal order while preserving the original textual structure.

2 Related Work

Many methods already exist for digitizing English documents, and they can serve as a useful starting point for building systems that handle various content. However, digitizing bilingual documents is much harder when the two languages use completely different writing patterns. Researchers have made great progress in detecting and recognizing clinical documents, which contain various languages, but no clinical documents dataset contains both low-resource and high-resource languages as shown in Table 1.

Table 1: Clinicians' Handwritten Text Dataset statistics

Ref	Study/Dataset	Language	Size
[8]	Persian clinical text dataset	Persian	188,963
[19]	Handwritten prescription dataset	English	24
[20]	Medical prescription dataset	English	40,000
[25]	Handwritten prescription images	English	893

2.1 Handwritten Clinical Prescriptions Dataset

Detecting and recognising highly illegible handwriting in clinical documents acquired under unconstrained conditions remains a central problem in computer vision. Accurate reading of handwritten text in these settings supports a wide range of practical applications. BHCP-DOC is the real-world clinical prescriptions dataset. This dataset is needed due to the unavailability of a large annotated dataset. Complementing these efforts, the BHCP-DOC dataset targets English and Urdu handwritten text specifically.

Table 2: Various Models Applied on Documents for Handwritten Text Detection[7], [26]

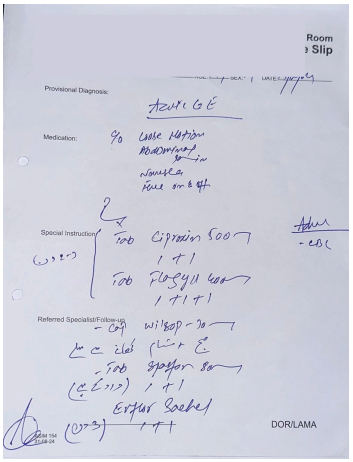
Model	Precision	Recall	F1
Yolo v3	0.857	0.811	0.833
Yolo v4	0.8515	0.8045	0.827
Yolo v5	0.897	0.874	0.885
Yolo v7	0.836	0.777	0.805
Faster R-CNN	0.930	0.8670	0.897

2.2 Real Bilingual Text Detection and Recognition

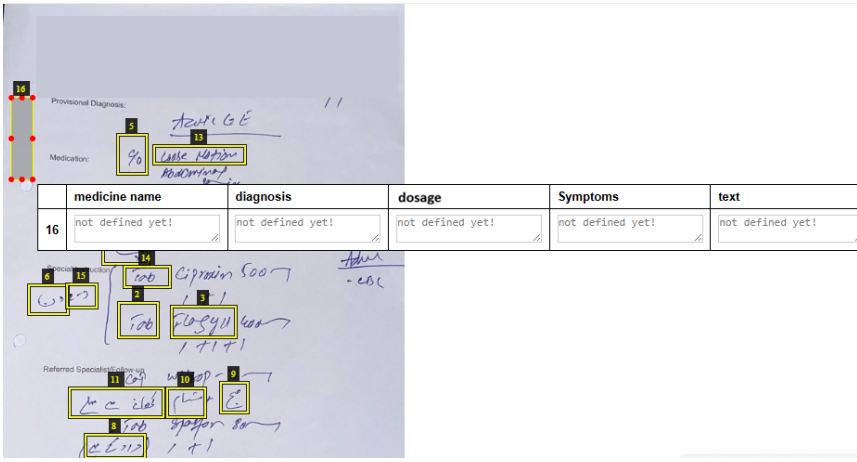
Text detection aims to localize text regions and is a critical prerequisite for recognition. Early approaches relied on hand-crafted cues such as edges, the stroke width transform (SWT), and maximally stable extremal regions (MSER) [10]. Traditional handcrafted approaches can be effective in controlled environments; their performance often deteriorates in real-world scenes characterized by cluttered backgrounds, uneven illumination, and varied typographic styles. Table 2 shows that Faster R-CNN achieves the highest F1 score among the listed methods, while YOLO variants provide competitive detection performance and are often preferred when faster inference is required [4].

Convolutional text detection models such as the Connectionist Text Proposal Network and the Efficient Accurate Scene Text Detector generally find text regions more consistently than traditional handcrafted pipelines. YOLO [12], as a detector, can locate text in one forward pass and uses global context to produce fast and accurate predictions. Building on this idea, several YOLO variants have been adapted for handwritten text detection [24], improving localization efficiency and capturing script-specific structure more effectively. With deep learning, models that combine convolutional encoders with recurrent decoders, such

as LSTMs, became a common choice for sequence modelling. CRNN, which uses a CNN to extract visual features, an RNN to capture sequential dependencies, and CTC to produce alignment-free transcriptions. The deep text recognition benchmark and ViTSTR improved accuracy by modelling long-range dependencies more effectively [1]. The ResNet50 CTC model uses a deep ResNet50 backbone to extract features and then applies recurrent or fully connected temporal modelling layers for transcription. This setup delivered the best performance, showing better robustness and generalization for multilingual and complex text line recognition [17].



(a) Bilingual Handwritten Clinical Prescription



(b) Prescription Annotation VIA VGG

Fig. 2: Anonymised Clinical Prescription and its Annotation

3 BHCP-DOC Dataset

We present a comprehensive dataset of bilingual handwritten clinical prescription texts in a real-world scenario, termed BHCP-DOC.

Prescriptions were collected from different clinicians by visiting different hospitals and captured using a mobile phone camera with a 5MP ultrawide lens and a 2MP macro camera. As illustrated in figure 2a, it shows challenging real-world writing styles, including unclear text, partial occlusion, unreadable, partially spelled words, perspective distortion, and diverse text orientations, scales, joined-up writing, and bilingual fonts. All prescription images were anonymised before release, and no personally identifiable patient, doctor, or hospital information is included in the public version. Data collection and release followed the applicable institutional data-handling requirements. The released dataset contains only anonymised prescriptions under GDPR regions required for OCR research. We reported the dataset statistics that characterize content diversity and difficulty. The BHCP-DOC dataset and the associated evaluation protocol

are publicly available online[5] to facilitate transparent benchmarking and reproducible research.

Table 3: Dataset statistics

Number of Doctors	205
Handwritten Text Lines	11,008
Training Lines	8,480
Validation Lines	1,264
Test Lines	1,264
English Lines	8,468
Urdu Lines	1,329
Mixed Lines	1,211

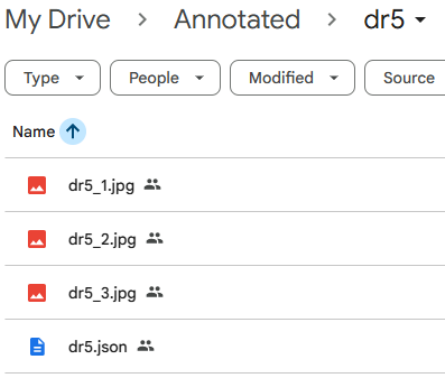


Fig. 3: Dataset Managing Protocol

Dataset Handling Table 3 presents the statistics of the clinical handwritten text dataset. The dataset consists of clinical documents collected from 205 doctors, each contributing a variable number of prescriptions. To ensure consistent organization and traceability, a unique ID was assigned to every doctor. The corresponding folder for each doctor was renamed using this ID, and the documents within each folder were labeled sequentially as ID_1, ID_2, as shown in Figure 3.

3.1 BHCP-DOC Annotation

After we created the dataset, classes for text localisation and transcription, the procedure used to validate annotation quality, and the strategy for splitting the corpus into training and validation sets were finalised. We defined five semantic classes to classify the text in prescription: medicine names, diagnosis, symptoms, text, and dosage. Annotations were created manually with the VGG Image Annotator (VIA), a web-based tool that supports different annotation shapes (polygon, square, circle, and rectangle) and attribute labeling (classes), helping ensure

[5] `https://l3i-share.univ-lr.fr/2025BHCP-DOC/`

Table 4: Word and character counts across the train, validation, and test splits.

Split	Urdu words	English words	Total words	Characters
Train	146,454	257,851	404,305	4,516,410
Validation	27,986	47,795	75,781	817,573
Test	32,737	58,629	91,366	1,022,550
Total	**207,177**	**364,275**	**571,452**	**6,356,533**

consistent marking of text instances relevant to their transcriptions [9]. We used a rectangular bounding box as shown in Figure 2b. The process starts by drawing a rectangular bounding box around each handwritten text. Rectangular bounding boxes are commonly used to annotate handwritten text because they provide a simple, consistent, and efficient way to mark text regions, as used in [2],[22], [6], [14]. Handwritten text can vary greatly in shape, size, style, and orientation; drawing precise boundaries is often difficult and subjective. Using rectangular boxes makes the annotation process easier, reduces differences between annotators, and still provides enough coverage of the text for detection and recognition tasks. All handwritten text in each document was manually annotated. Native Urdu speakers carried out the annotation of the Urdu text, and the first author performed additional verification when required. Once the annotation was completed, an expert clinician reviewed the labeled text to confirm the accuracy of all classes, given the sensitive nature of clinical documents and the importance of minimizing annotation errors. Each text receives exactly one label, which ensures consistent ground truth for detection, recognition, and downstream information extraction. Annotations are stored per doctor in JSON. These structured JSON entries facilitate reproducible preprocessing and provide a clean interface for training and evaluating deep models for bilingual handwritten clinical text.

3.2 BHCP-DOC Statistics

There are a total of 893 real-world handwritten clinical prescriptions collected from various hospitals and clinicians. These clinical prescriptions are bilingual, containing challenging English and Urdu handwritten text. There are a total of 11,008 handwritten text lines. To support model training and evaluation, a stratified split is applied to the dataset, and there are 8,480 training lines, 1,264 validation lines, and 1,264 test lines. It consists of bilingual text, with 8,468 English lines and 1,329 Urdu lines. The remaining 1,211 lines are grouped as mixed/other and include numeric-only, symbol-only, or illegible handwritten prescription text. These numbers highlight the size and bilingual nature of the handwritten clinical documents dataset. BHCP-DOC comprises 571,452 handwritten Urdu and English word tokens and 6,356,533 characters. To characterize linguistic coverage, we analyze the most frequent characters and words and report their distributions across the training, validation, and test splits. Table 4 summarizes the word and character counts for each split. Figures 4, 5, and 6 show the distributions of the most frequent words and characters under the selected tokenisation tools. These distributions show that BHCP-DOC contains

substantial coverage of high-frequency characters and words, supporting stable optimization of recognition models.

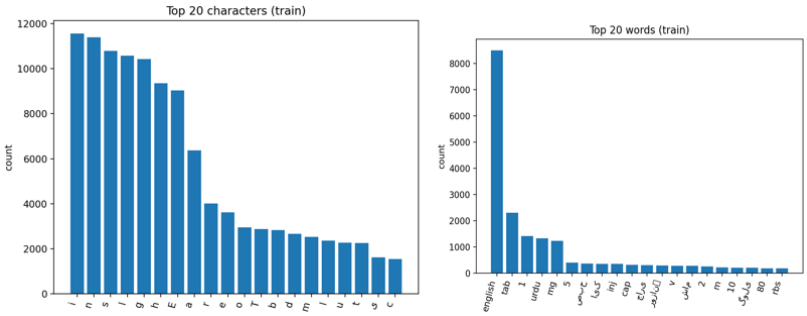


Fig. 4: Most frequent words and characters in the training split using UrduHack

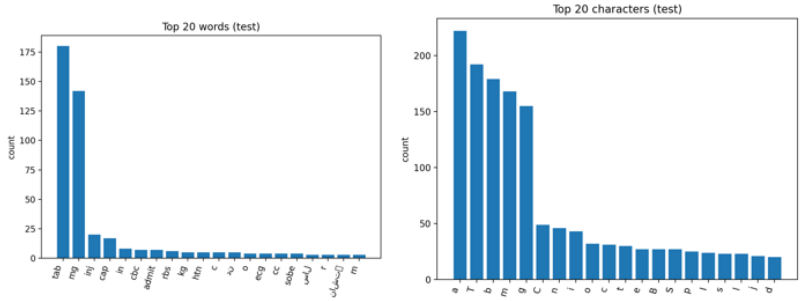


Fig. 5: Most frequent words and characters in the test split using Stanza.

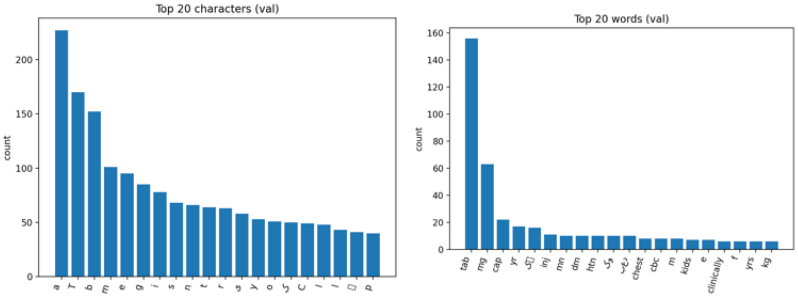


Fig. 6: Most frequent words and characters in the validation split using NLTK.

4 Evaluation Metrics

To measure how accurately deep learning models recover text sequences from images, most studies use the Levenshtein distance. For BHCP-DOC, however, a plain CER calculation can be misleading because the script and writing style introduce additional complexity. To evaluate recognition more fairly in this setting, we applied a character normalization strategy similar to the one described

in [3] and also used in [5]. The idea is to break each character cluster into its basic components, apply rule-based corrections at that level, and then rebuild the cluster in a consistent canonical form. The process starts by removing zero-width elements and identifying the base character of each syllable, either a consonant or an independent vowel. We compute CER using Levenshtein distance, following prior work [21], as shown in Equation 1:

$$CER = \frac{S + I + D}{N} \tag{1}$$

where S, I, and D denote substitutions, insertions, and deletions, respectively, and N is the number of characters in the ground-truth transcription. Before computing CER and WER, both predictions and ground-truth transcriptions are processed using the same Urdu-English normalization pipeline. The pipeline removes zero-width characters, normalizes character-cluster order, and standardizes minor script-level variations. CER is then computed at the character level using Levenshtein distance, while WER is computed after tokenising the normalised prediction and ground truth into word units. A comparison with standard tools such as Jiwer and TorchMetrics shows that our normalized metric produces slightly lower CER and WER values for cases involving minor vowel placement or character-order variations.

5 Benchmark Model and Performance

Our benchmark framework decomposes bilingual clinical prescription OCR into two main tasks: text-line detection and text-line recognition, as shown in Figure 7. Detection localizes handwritten text regions using detectors such as YOLO variants, Faster R-CNN, and EAST. Recognition maps cropped text-line images to character sequences using CRNN-based models, attention-based decoding, and TrOCR.

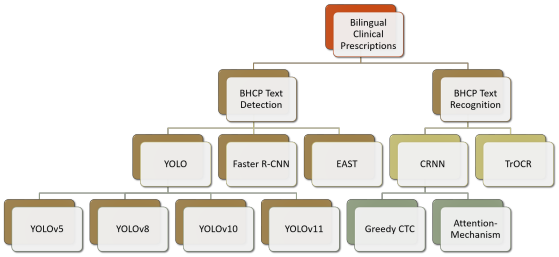


Fig. 7: Bilingual Handwritten Text Detection and Recognition Pipeline

5.1 BHCP-DOC Text Line Detection

Detecting doctors' handwritten text is challenging due to complex character shapes and messy handwriting. We address these factors with single-stage YOLO

detectors. Prior work applied YOLOv8s, YOLOv10s, and YOLOv11n to Khmer Scene Text character and line detection and evaluated several YOLO variants for Khmer text lines [16]. Building on these findings, we compare YOLOv5nu, YOLOv8n, YOLOv10s, and YOLOv11n for Urdu-English line detection in table 5. Figure 8 shows text-line detection results using YOLOv5nu and YOLOv8n, while Figure 9 shows results using YOLOv10s and YOLOv11n on clinical images.

We train recent YOLO families for BHCP text-line detection. YOLOv8n uses anchor-free heads with stronger data augmentation, which can improve localization. YOLOv10s and YOLOv11n target higher efficiency while keeping performance on dense or small targets. Each image is sized to 640 pixels. Labels follow the standard YOLO format ((x_center, y_center, width, height, class_id)) with a single class, where class id 0 denotes a BHCP-DOC text line [15]. We report precision, recall, and mAP at 0.50 IoU, and mAP averaged from 0.50 to 0.90 IoU in steps of 0.05. Table 5 summarizes the results YOLOv11n attains the highest precision at 0.750 and the highest mAP@50 at 0.7286. YOLOv10s achieves the best recall at 0.684, and the highest mAP averaged across 0.50 to 0.90 IoU at 0.427, indicating stronger consistency under stricter overlap thresholds. YOLOv8n provides a balanced outcome with precision 0.634, recall 0.608, mAP@50 0.6558, and mAP@50–90 0.349. The YOLOv5nu baseline is lower in precision at 0.608 and mAP@50 at 0.6419, with recall 0.6228 and mAP@50–90 0.334. YOLOv11n is preferable when precision and mAP@50 are the primary objectives, while YOLOv10s is better when higher recall and stability across IoU thresholds are more critical. YOLOv8n offers a compact alternative with competitive accuracy, and YOLOv5nu serves as a reference baseline for this task.

Table 5: Various YOLO Models Applied on Clinical Documents for Detection of Handwritten Text

Model	**Recall**	**Precision**	**mAP@50**	**mAP@50-90**
YOLOv5nu	0.6228	0.608	0.6419	0.334
YOLOv8n	0.608	0.634	0.6558	0.349
YOLOv10s	0.684	0.657	0.7185	0.427
YOLOv11n	0.657	0.750	0.72858	0.40

(a) Detecting Results from YOLOv5nu, text line is detecting in each line

(b) Detecting Results from YOLOv8n, text line is detecting in each line

Fig. 8: Text line Detection using YOLOv5nu and YOLOv8n models

(c)Detecting Results from YOLOv10s, text line is detecting in each line

(d)Detecting Results from YOLOv11n, text line is detecting in each line

Fig. 9: Text line Detection using YOLOv10s and YOLOv11n models

5.2 Real BHCP-DOC Text Line Recognition

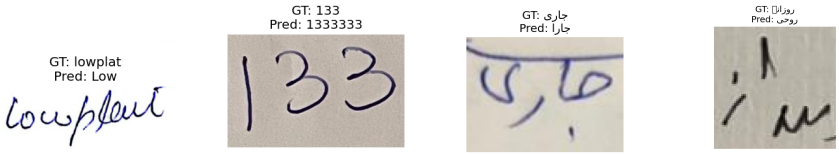
Given a ground-truth cropped region containing a single text line, the recognition task aims to accurately extract the text. To assess the effectiveness of different methods, we evaluated several state-of-the-art models, including CRNN-based models and TrOCR. The first model combines a CRNN with VGG feature extractors and a greedy CTC decoder. This architecture provides a simple and fast recognition baseline by extracting visual features from the input line image and mapping the resulting sequence to output characters using CTC decoding. However, this model obtains the highest error rates, with a CER of 1.0 and a WER of 1.28. These results indicate that the greedy CTC-based CRNN struggles to handle the irregular, noisy, and bilingual handwritten text present in BHCP-DOC.

The second model uses a CRNN with a BiLSTM backbone and an attention-based Seq2Seq decoder. The BiLSTM captures contextual dependencies in both forward and backward directions, while the attention decoder dynamically focuses on relevant regions of the encoded feature sequence during transcription. This model achieves a CER of 0.524 and a WER of 0.585, showing a clear improvement over the greedy CTC-based CRNN in table 6. These results suggest that attention-based decoding is more suitable than greedy CTC decoding for handling complex bilingual prescription handwriting. The best recognition per-

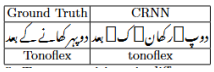
Table 6: Recognition performance of different models on bilingual handwritten clinical text.

Model	Backbone	Decoder	CER	WER
CRNN	VGG	Greedy CTC decoder	1.0	1.28
CRNN	BiLSTM	Attention-based Decoder (Seq2Seq)	0.524	0.585
TrOCR	BEiT (ViT-style)	RoBERTa-based Transformer Decoder	0.448	0.567

formance is obtained by the TrOCR-based model, which uses a BEiT-style vision encoder and a RoBERTa-based transformer decoder. TrOCR achieves the lowest CER of 0.448 and the lowest WER of 0.567 among the evaluated models. This indicates that transformer-based encoder-decoder recognition benefits from pretraining and stronger sequence modeling, making it more effective for noisy Urdu-English handwritten clinical prescriptions. The remaining error rates show that BHCP-DOC is still a challenging benchmark due to irregular writing, overlapping strokes, weak word boundaries, and mixed-script content, as illustrated in Figure 10.



(a) English Text Recognition (b) Numeric Text Recognition (c) Urdu Text Recognition (d) Urdu Text Recognition



(e) Text Recognition

Fig. 10: Handwritten Text Recognition of Clinical Documents via different models

5.3 Discussions and Limitations

Detection. Among the evaluated detectors, YOLOv11n achieved the highest precision of 0.750 and the highest mAP@50 of 0.7286, while YOLOv10s achieved the highest recall of 0.684 and the highest mAP@50–90 of 0.427. This indicates that YOLOv11n is preferable when reducing false positives is important, whereas YOLOv10s is better when recall and stricter localisation quality are prioritised. The detector shows clear limitations with bilingual text in real-world prescription images, as illustrated in Figure 11, where detections are inconsistent, and some instances are missed. These failure modes indicate a need for improved detection strategies that remain reliable for low-resolution, messy handwriting, unreadable words, and low-contrast text in complex real-world scenes.

Fig. 11: Undetected Handwritten Text of Clinical Documents Using YOLO

Recognition. For recognition, TrOCR achieved the best performance among the evaluated models, with a CER of 0.448 and a WER of 0.567, as shown in Table 6. The CRNN model with a BiLSTM backbone and attention-based Seq2Seq decoder ranked second, obtaining a CER of 0.524 and a WER of 0.585. In contrast, the CRNN model with VGG features and greedy CTC decoding produced the highest error rates, with a CER of 1.0 and a WER of 1.28. These results suggest that greedy CTC decoding is not sufficient for the severe variability present in bilingual handwritten clinical prescriptions, while attention-based

and transformer-based models provide stronger sequence modeling. TrOCR performs best because its encoder-decoder transformer architecture can better capture long-range dependencies and contextual information in noisy Urdu-English handwritten text. However, the remaining recognition errors show that BHCP-DOC remains challenging, especially for illegible medicine names, overlapping strokes, inconsistent baselines, numeric expressions, and mixed-script text. As shown in Figure 10, several recognition errors involve uneven pen strokes, overlapping characters, unclear word boundaries, and numeric expressions, and inconsistent baselines. These cases highlight the need for larger and more varied training data, stronger preprocessing, script-aware tokenisation, and domain-specific language modeling for clinical prescription OCR.

6 Conclusion

This work focuses on detecting and recognising bilingual handwritten text in real clinical prescriptions. We curated a large and diverse dataset with 11,008 text lines taken from 893 prescriptions, each annotated with rectangular bounding boxes and line-level transcriptions. The collection is designed to reflect real-world conditions, including decorative or stylised writing, noisy and unclear text, hard-to-read medicine names, messy character shapes, blur, low lighting, curved baselines, cluttered backgrounds, and occlusions. As a result, it provides a demanding and realistic benchmark for both training and evaluation. We also report benchmarks across state-of-the-art detectors and recognisers. For detection, YOLOv11n achieves the highest precision and mAP@50, while YOLOv10s achieves the highest recall and mAP@50–90. For recognition, TrOCR achieves the lowest CER and WER, outperforming the CRNN-based baselines. These results provide reproducible baselines and show that BHCP-DOC remains challenging for both text localisation and mixed-script handwritten text recognition. Together, the dataset, annotations, evaluation protocol, and benchmark results support further research on ambiguous characters, complex layouts, and low-resolution handwritten clinical text.

7 Future Work

We will extend BHCP-DOC from a line-level OCR benchmark into a scalable bilingual prescription understanding testbed by (i) increasing the number of real prescriptions and writer diversity, and (ii) introducing evaluation protocols that explicitly measure cross-writer and cross-site generalization under realistic acquisition noise (blur, low illumination, clutter, and severe stroke degradation). Building on the current annotations and baseline pipelines, we will refine supervision with script-aware normalization and difficulty tags to support principled analysis of failure modes in mixed Urdu–English handwriting, including uncertainty-aware scoring for partially ambiguous content.
We will study end-to-end and multi-task architectures that jointly optimize detection and recognition, reducing error propagation common in two-stage

pipelines and enabling global context utilization across the prescription image. We will also explore data-centric and representation-learning strategies for low-resource Urdu handwriting, including self-supervised and semi-supervised pre-training on unlabeled clinical scans, synthetic bilingual augmentation with controllable noise, and lexicon- or constraint-guided decoding for domain-specific medical terms. We will then expand supervision toward clinically grounded categories to support downstream tasks such as bilingual text-line classification, key information extraction, and medical text normalization from noisy handwritten inputs.

Acknowledgments This work was supported by the "PHC PERIDOT" program (project number:51405PL), funded by the French Ministry for Europe and Foreign Affairs, the French Ministry for Higher Education and Research, and the Higher Education Commission (HEC) in Pakistan.

References

1. Afkari-Fahandari, A., Shabaninia, E., Asadi-Zeydabadi, F., Nezamabadi-Pour, H.: A comprehensive survey of transformers in text recognition: Techniques, challenges, and future directions. ACM Computing Surveys (2025)
2. Al-Barhamtoshy, H.M., Jambi, K.M., Rashwan, M.A., Abdou, S.M.: An arabic manuscript regions detection, recognition and its applications for ocring. Transactions on Asian and Low-Resource Language Information Processing **22**(1), 1–28 (2023)
3. Arnold, M.: Multilingual research projects: Non-latin script challenges for making use of standards, authority files, and character recognition. Digital Studies/Le champ numérique **12**(1), 1–36 (2022)
4. Bilous, N., Malko, V., Frohme, M., Nechyporenko, A.: Comparison of cnn-based architectures for detection of different object classes. AI **5**(4), 2300–2320 (2024)
5. Buoy, R., Iwamura, M., Srun, S., Kise, K.: Toward a low-resource non-latin-complete baseline: an exploration of khmer optical character recognition. IEEE Access **11**, 128044–128060 (2023)
6. Chandio, A.A., Asikuzzaman, M., Pickering, M.R., Leghari, M.: Cursive text recognition in natural scene images using deep convolutional recurrent neural network. IEEE Access **10**, 10062–10078 (2022)
7. Cheema, M.D.A., Shaiq, M.D., Mirza, F., Kamal, A., Naeem, M.A.: Adapting multilingual vision language transformers for low-resource urdu optical character recognition (ocr). PeerJ Computer Science **10**, e1964 (2024)
8. Dashti, S.M.S., Dashti, S.F.: Improving the quality of persian clinical text with a novel spelling correction system. BMC Medical Informatics and Decision Making **24**(1), 220 (2024)
9. Dutta, A., Zisserman, A.: The via annotation software for images, audio and video. In: Proceedings of the 27th ACM international conference on multimedia. pp. 2276–2279 (2019)
10. Dutta, K., Sarkhel, R., Kundu, M., Nasipuri, M., Das, N.: Natural scene text localization and detection using mser and its variants: a comprehensive survey. Multimedia Tools and Applications **83**(18), 55773–55810 (2024)
11. Fateh, A., Fateh, M., Abolghasemi, V.: Multilingual handwritten numeral recognition using a robust deep network joint with transfer learning. Information Sciences **581**, 479–494 (2021)

12. Gallagher, J.E., Oughton, E.J.: Surveying you only look once (yolo) multispectral object detection advancements, applications and challenges. IEEE Access (2025)
13. Kubra, K.T., Umair, M., Zubair, M., Naseem, M.T., Lee, C.S.: Detection and recognition of bilingual urdu and english text in natural scene images using a convolutional neural network–recurrent neural network combination with a connectionist temporal classification decoder. Sensors **25**(16), 5133 (2025)
14. Mahadevkar, S., Patil, S., Kotecha, K., Abraham, A.: A comparison of deep transfer learning backbone architecture techniques for printed text detection of different font styles from unstructured documents. Peerj Computer Science **10**, e1769 (2024)
15. Nom, V., Bakkali, S., Luqman, M.M., Coustaty, M., Ogier, J.M.: Khmerst: A low-resource khmer scene text detection and recognition benchmark. In: Proceedings of the Asian Conference on Computer Vision. pp. 1777–1792 (2024)
16. Nom, V., Keo, S., Bakkali, S., Luqman, M.M., Coustaty, M., Rossinyol, M., Ogier, J.M.: Wildkhmerst: A comprehensive dataset and benchmark for khmer scene text detection and recognition in the wild. In: International Conference on Document Analysis and Recognition. pp. 351–368. Springer (2025)
17. Nugraha, G.S., Darmawan, M.I., Dwiyansaputra, R.: Comparison of cnn's architecture googlenet, alexnet, vgg-16, lenet-5, resnet-50 in arabic handwriting pattern recognition. Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control (2023)
18. Phung, T.M., Dinh, M.N., Dang, D.P.T., Van, H.M.T., Thwaites, C.L.: A machine learning-based approach to vietnamese handwritten medical record recognition (2020)
19. Rani, S., Rehman, A.U., Yousaf, B., Rauf, H.T., Nasr, E.A., Kadry, S.: Recognition of handwritten medical prescription using signature verification techniques. Computational and Mathematical Methods in Medicine **2022**(1), 9297548 (2022)
20. Razdan, A., Raghavan, A., Panandikar, M., Pol, A., Hedaoo, K., Patil, R.: Recognition of handwritten medical prescription using cnn bi-lstm with lexicon search. In: 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT). pp. 1–6. IEEE (2023)
21. ul Sehr Zia, N., Naeem, M.F., Raza, S.M.K., Khan, M.M., Ul-Hasan, A., Shafait, F.: A convolutional recursive deep architecture for unconstrained urdu handwriting recognition. Neural Computing and Applications **34**(2), 1635–1648 (2022)
22. Singh, S., Garg, N.K., Kumar, M.: Vgg16: Offline handwritten devanagari word recognition using transfer learning. Multimedia Tools and Applications **83**(29), 72561–72594 (2024)
23. Tabassum, S., Abedin, N., Rahman, M.M., Rahman, M.M., Ahmed, M.T., Islam, R., Ahmed, A.: An online cursive handwritten medical words recognition system for busy doctors in developing countries for ensuring efficient healthcare service delivery. Scientific reports **12**(1), 3601 (2022)
24. Turki, H., Elleuch, M., Kherallah, M.: Multi-lingual scene text detection containing the arabic scripts using an optimal then enhanced yolo model. In: International Conference on Model and Data Engineering. pp. 47–61. Springer (2023)
25. Zaman, M.F.U., Rahman, B., Rubyat, A., Halder, N., Mahmud, A., Islam, A., Amin, M.A.: Recognizing medicine names from bangladeshi handwritten prescription images using trocr. In: 2024 7th Asia Conference on Cognitive Engineering and Intelligent lnteraction (CEII). pp. 55–59. IEEE Computer Society (2024)
26. Zhong, Z., Sun, L., Huo, Q.: Improved localization accuracy by locnet for faster r-cnn based text detection. In: 2017 14th IAPR international conference on document analysis and recognition (ICDAR). vol. 1, pp. 923–928. IEEE (2017)