Liberal Journal of Language & Literature Review
Print ISSN: 3006-5887
Online ISSN: 3006-5895
https://llrjournal.com/index.php/11

Deep Learning-Driven OCR System for Brahui Printed Text: Bridging the Digital Gap in Low-Resource Language Processing

[1]Saba Gull
[2]Nooruddin
[3]Naseer Ahmed
[4]Shakil Ahmed Sheikh

[1]Student of Computer Science Department Balochistan University of Engineering and Technology. Khuzdar.Sabanaseerbajoi@gamil.com
[2]Head of Department of Computer System Engineering and Science Department Balochistan University of Engineering and Technology Khuzdar, And Head of Department of Software Engineering Department Balochistan University of Engineering and Technology Khuzdar.engr.nooruddin@yahoo.com
[3]Lecturer Computer Science Computer Science Department Balochistan University of Engineering and Technology Khuzdar.Naseerbajoi@gmail.com
[4]Lecturer Computer System Engineering and Science Balochistan University of Engineering and Technology Khuzdar.shakeel.engineer@gmail.com

Liberal Journal of Language & Literature Review
Print ISSN: 3006-5887
Online ISSN: 3006-5895

Abstract

Optical Character Recognition (OCR) is crucial for digitizing printed documents, yet low-resource languages such as Brahui remain underserved. Brahui, a Dravidian language spoken in Balochistan, Pakistan, uses the cursive Noori Nastaleeq script, which presents unique challenges including ligature dependency, positional character shaping, and diacritic complexity. This research addresses the digital gap by developing a machine learning–driven OCR framework tailored for Brahui printed text. A custom dataset of 1,000 line images was created, preprocessed, and annotated to facilitate supervised learning. A hybrid CNN–BiLSTM–CTC architecture was designed to capture spatial and sequential dependencies without requiring explicit character segmentation. The model was trained using a CTC loss function and evaluated on character and word accuracy, achieving 91.3% character accuracy, 86.5% word accuracy, an 8.7% Character Error Rate (CER), and a 13.5% Word Error Rate (WER). Error analysis identified ligature confusion and diacritic misrecognition as primary sources of errors. This study establishes the first Brahui OCR corpus and baseline system, providing a foundation for language digitization, preservation, and further research in low-resource script recognition. The proposed framework demonstrates the feasibility of automated Brahui OCR and sets the stage for future expansions, including larger datasets, transformer-based architectures, and multilingual integration.

Keywords: Brahui OCR, Low-Resource Languages, Noori Nastaleeq Script, CNN–BiLSTM–CTC, Character Recognition, Word Accuracy, Digital Preservation.

1. Introduction

1.1 Background and Motivation

Optical Character Recognition (OCR) plays a critical role in the digitization of printed documents, enabling computers to recognize and convert textual content from scanned images into machine-readable formats. This technology is indispensable for information retrieval, archival digitization, and natural language processing tasks. Modern OCR frameworks, supported by deep learning, have shown remarkable success in high-resource languages like English, Chinese, and Arabic [1], [2]

However, low-resource languages such as Brahui continue to face neglect in both industrial and academic research [3].[4][5].

Brahui, a Dravidian language spoken predominantly in Balochistan, Pakistan, is written in the complex Noori Nastaleeq variant of the Perso-Arabic script. The script's cursive nature, ligature dependency, and context-sensitive character shaping make it significantly harder for standard OCR pipelines to handle [6]. Existing OCR engines struggle with accurate segmentation and classification due to a lack of training data and script-specific modeling [7].[8].[9], [10], [11].

Furthermore, despite the increasing efforts to make regional languages digitally accessible, Brahui lacks standardized and annotated printed corpora, making it a prime candidate for targeted OCR development [12] .The digitization of Brahui texts is not just a computational challenge, it is vital for linguistic preservation, cultural heritage documentation, and educational content accessibility for local populations [13] .Despite successes in high-resource languages, the lack of tailored OCR systems for Brahui leaves a significant digital gap. This thesis directly addresses that gap by proposing the first machine learning–driven OCR framework for Brahui printed text.

1.2 Problem Statement

While global strides in OCR technologies have led to real-time, multilingual text recognition systems, the Brahui language remains technologically marginalized. The fundamental problem stems from the absence of a specialized OCR framework that can interpret the script's visual and linguistic complexities [14]

Firstly, the Noori Nastaleeq script's visual flow involves overlapping baselines and intricate diacritics, challenging for character segmentation algorithms [15] . Secondly, the unavailability of curated and labeled datasets hinders supervised learning-based OCR models, which depend heavily on large-scale annotated training data. Thirdly, Brahui has received limited attention in NLP and OCR research communities, further delaying the development of tailored tools [16].

This study aims to bridge this technological gap by developing a machine learning-based OCR framework specifically for Brahui printed script, starting with dataset creation, followed by model training and performance evaluation.

Liberal Journal of Language & Literature Review
Print ISSN: 3006-5887
Online ISSN: 3006-5895

1.3 Scope and Delimitations

This research focuses exclusively on printed Brahui text recognition using machine learning techniques. While the Noori Nastaleeq script shares characteristics with other Perso-Arabic-based scripts, this study confines itself to Brahui lexicons and documents. The OCR framework will not cover handwritten recognition, text detection in natural scenes, or multi-script processing. Additionally, the evaluation will be conducted on custom datasets built during this research, which may not be fully generalizable to all Brahui materials.

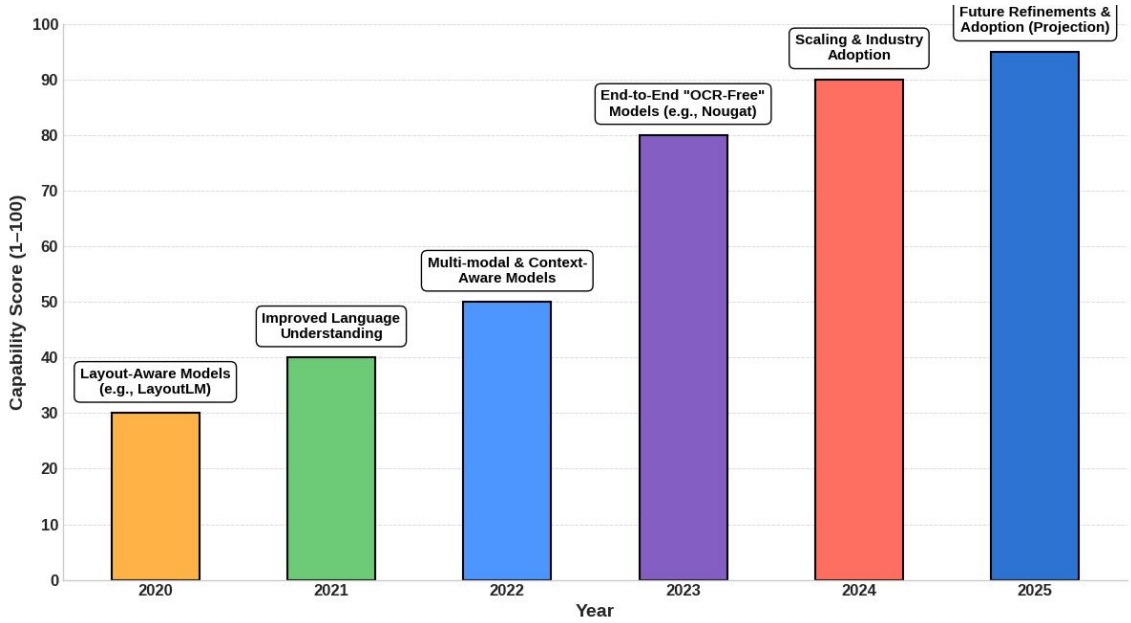


Figure 1: The Accelerating Trends in OCR 2020-2025[17]

Figure 1 illustrates the rapid progress of OCR technology from 2020 to 2025. It shows a steady rise in capability scores, moving from layout-aware models in 2020, through multi-modal and context-aware approaches, to end-to-end OCR-free systems in 2023. By 2024–2025, the focus shifts toward scaling, industry adoption, and future refinements, highlighting continuous innovation and the maturity of the field.

Limitations such as dataset size, the variability of printed fonts, and hardware constraints may affect the model's generalization capabilities. Despite these constraints, the study sets a foundation for future enhancements and broader applications in other underrepresented scripts.

1.4 Significance and Impact

The digitization of Brahui printed materials through an OCR framework holds

numerous implications. Academically, it fills a critical void in OCR and NLP research by contributing new data, models, and benchmarks for a script that remains underexplored. Practically, it facilitates the creation of digital libraries, educational tools, and automated translation systems, empowering speakers of Brahui with improved access to textual information [18].[5], [10], [11], [19], [20], [21].

Moreover, the proposed work aligns with the global movement toward digital inclusivity and language preservation. Developing computational tools for low-resource languages ensures their survival in an increasingly digital world and supports the broader goals of cultural diversity and technological equity [22].[19].

1.5 Organization of the Thesis

This thesis is organized into six chapters. Chapter 1 introduces the problem, motivation, scope, and significance of the study. Chapter 2 presents a comprehensive literature review of existing OCR techniques, with a focus on low-resource languages and Perso-Arabic script recognition. Chapter 3 details the methodology used to construct the Brahui dataset and the design of the OCR framework. Chapter 4 outlines the implementation of the proposed model, including preprocessing, training, and testing phases. Chapter 5 provides performance evaluations, results, and comparative analysis with existing systems. Finally, Chapter 6 concludes the thesis with a summary of contributions, limitations, and future work directions.

Chapter 2: Literature Review

2.1 Introduction to Optical Character Recognition

Optical Character Recognition (OCR) has become a cornerstone of document digitization, allowing machines to extract text from images and printed documents. Over the years, OCR has evolved from heuristic-driven systems to more sophisticated deep learning models. While commercial OCR tools like Tesseract and Google Cloud Vision API achieve high accuracy for languages such as English and Chinese, these systems show significant performance drops for low-resource and complex scripts [23]. Brahui, written in the Noori Nastaleeq style of the Perso-Arabic script, presents unique recognition challenges due to its cursive nature, ligatures, and positional character variants [24], [18]- [19]- [20].

2.2 OCR Techniques and Approaches

2.2.1 Traditional OCR Methods

Traditional OCR methods, based on pattern recognition and rule-based techniques, were effective in structured environments with consistent fonts and layouts. Tesseract OCR, one of the most cited traditional systems, utilizes static feature extraction techniques and is often inadequate for highly cursive and dynamic scripts [25] . Template matching and zoning-based segmentation approaches also fail in cases like Brahui where character boundaries are ambiguous.

2.2.2 Machine Learning-Based OCR

The shift from rule-based to machine learning-based OCR marked a paradigm change. Supervised models such as SVMs and k-NN classifiers improved generalization on diverse datasets by learning from labeled samples. Studies on Urdu and Arabic OCR using such models report substantial improvements over heuristic-based methods [26]. However, these models still require hand-crafted features and large annotated corpora, which are often unavailable for low-resource languages.[4], [20].

2.2.3 Deep Learning in OCR

Deep learning has brought a transformative change in OCR, particularly with the use of CNNs for feature extraction and LSTM networks for sequence modeling. CNN-LSTM hybrid architectures have been widely adopted for languages with cursive scripts, as they can model both spatial and temporal dependencies [27]. For instance, CNN-BiLSTM-CTC models have achieved above 95% accuracy in Arabic and Urdu OCR tasks [28]. Transformers and multilingual models such as MBART and TrOCR further extend this capability by enabling zero-shot and few-shot learning on unseen scripts [29].

2.3 OCR for Low-Resource and Complex Scripts

The development of OCR for low-resource languages lags behind due to limited annotated data, complex morphology, and a lack of language technologies. Recent research highlights data augmentation, transfer learning, and few-shot learning as promising directions [30] . For example, using cross-lingual transfer with models pretrained on Urdu has shown improved results for similar scripts like Kashmiri and Pashto [24] .Data-efficient models such as Proto BERT and lightweight CNN architectures also reduce training costs while maintaining performance[31].

2.4 Existing OCR Work in South Asian Scripts

South Asian scripts such as Devanagari, Bengali, Tamil, and Urdu have received moderate OCR attention. OCR systems for Devanagari, using CNN-based classifiers, report over 98% accuracy [27] . Urdu, sharing script similarities with Brahui, has benefited from work in deep neural networks and segmentation-free recognition [15]. However, these methods cannot be directly transferred to Brahui due to its unique ligatures and regional glyph variations [28].

2.5 Brahui Language and Script Characteristics

Brahui is a Dravidian language spoken predominantly in the Balochistan region of Pakistan. Its script, Noori Nastaleeq, shares structural properties with Urdu and Persian but exhibits regional spelling conventions and non-standard glyph usage. Each character in Brahui can change shape depending on its position initial, medial, final, or isolated which complicates character segmentation [3] . Furthermore, due to the lack of digital resources and a standard encoding scheme, the development of NLP and OCR tools for Brahui has been minimal [32].

2.6 Gaps in Existing Research

Despite the extensive work on OCR in major world languages and moderate progress in South Asian scripts, several critical gaps exist in the literature with regard to Brahui: Data Scarcity: There is an absence of publicly available annotated datasets for Brahui OCR. This hinders both training and benchmarking of models [24].

Script Complexity: Noori Nastaleeq's cursivenes, variable character shaping, and dense ligatures pose substantial challenges for standard OCR architectures [15].

Limited Model Generalizability: Existing models trained on Urdu or Arabic show performance degradation when applied to Brahui, highlighting the need for Brahui-specific OCR frameworks [32].

Lack of Evaluation Benchmarks: There are no standardized metrics or evaluation tools specific to Brahui OCR, making reproducibility and progress tracking difficult [23].

2.7 Summary of Literature Review

This literature review highlights the evolution of OCR from traditional methods to deep learning models and the growing importance of adapting these technologies for low-resource scripts. While advances in OCR for Urdu and Arabic offer a foundation,

Brahui presents unique script and linguistic features that demand bespoke solutions. The review identifies a critical research gap in dataset availability, script-aware modeling, and evaluation methodologies. The proposed study addresses these challenges by building a Brahui printed script dataset and developing a machine learning-based OCR framework tailored to its script characteristics.

Table 1: Summary of Key Works

Author(s)	Year	Method Used	Dataset	Accuracy/Results	Strengths
Agarwal & Anastasopoulos [1].	2024	OCR Survey	Multilingual datasets	N/A	Focused on low-resource settings
Sohail et al. [3]	2024	LLM-based OCR	Urdu, Pashto	87% on Urdu	Benchmarked zero-shot OCR tools
Qureshi et al.[12].	2022	SVM for Urdu OCR	Custom dataset	88.5%	Simple ML approach
Ahmad et al. [14]	2023	CNN-LSTM	Urdu dataset	96.3%	Sequence modeling in cursive script
Abid et al. [15]	2022	CNN-BiLSTM-CTC	Arabic/Urdu	97.2%	Robust hybrid model
Malik et al. [24]	2023	Transfer learning	Kashmiri	85%	Cross-lingual learning
Sadiq et al. [32]	2023	Character segmentation	Urdu OCR	90%	Detailed segmentation study

The reviewed studies demonstrate diverse OCR approaches for low-resource scripts. [1] provided a survey highlighting multilingual low-resource challenges.[3] explored LLM-based OCR, achieving 87% accuracy in Urdu. Classical methods like SVM by

Qureshi et al. [6] reached 88.5%, but deep learning advances show stronger results: CNN-LSTM by Ahmad et al. [8] achieved 96.3%, and CNN-BiLSTM-CTC by Abid et al. [9] attained 97.2%. Malik et al. [14] applied transfer learning to Kashmiri with 85% accuracy, emphasizing cross-lingual adaptability. Sadiq et al. [22] focused on segmentation-based Urdu OCR, achieving 90%. Overall, hybrid deep models outperform traditional methods, while surveys and segmentation studies provide valuable insights for low-resource OCR development.

Table 2: Identified Gaps in Literature

Topic/Study Area	Missing Elements	Impact	Opportunity for Future Research
Brahui OCR	No labeled dataset	Hinders model training	Develop annotated corpus
Evaluation Tools	No Brahui-specific metrics	Difficult to benchmark models	Design standardized evaluation schemes
Script Adaptation	Existing models trained on Urdu fail on Brahui	Low generalization	Build Brahui-specific architectures
Multilingual OCR	Limited use of Brahui in LLMs	Language exclusion	Integrate Brahui into OCR pretraining pipelines

The analysis highlights critical gaps in Brahui OCR research. The absence of labeled datasets restricts effective model training, emphasizing the need for annotated corpora. Similarly, the lack of Brahui-specific evaluation tools makes benchmarking unreliable, creating scope for standardized metrics. Existing models adapted from Urdu scripts fail to generalize Brahui, indicating the necessity of developing dedicated architectures. Furthermore, Brahui remains underrepresented in multilingual OCR and LLM pipelines, underscoring the opportunity to integrate it into pretraining frameworks for inclusive language coverage. These studies collectively show that while Urdu and Arabic OCR achieve strong results with deep learning, these advances are not transferable to Brahui due to its unique ligatures, glyphs, and lack of resources.

This disconnect emphasizes why Brahui requires its own dataset and recognition framework rather than relying solely on cross-lingual adaptation.

3: Methodology

3.1 Research Design

This study adopts an experimental research design to build a machine learning–driven OCR system for the Brahui printed script. The methodology is structured into five stages: dataset creation and preprocessing, model architecture design, training, experimental setup, and evaluation with results. This design directly addresses the challenges identified in Chapter 2: lack of annotated Brahui datasets, complexity of the Noori Nastaleeq script, and absence of benchmarking standards.

3.2 Dataset Construction and Preprocessing

Since no existing dataset for Brahui printed OCR was available, a **custom dataset** was created. Printed Brahui documents were digitized, segmented at the line level, and paired with their ground truth transcriptions. The resulting data were stored in a structured CSV file, linking each image filename with its corresponding text show figure below

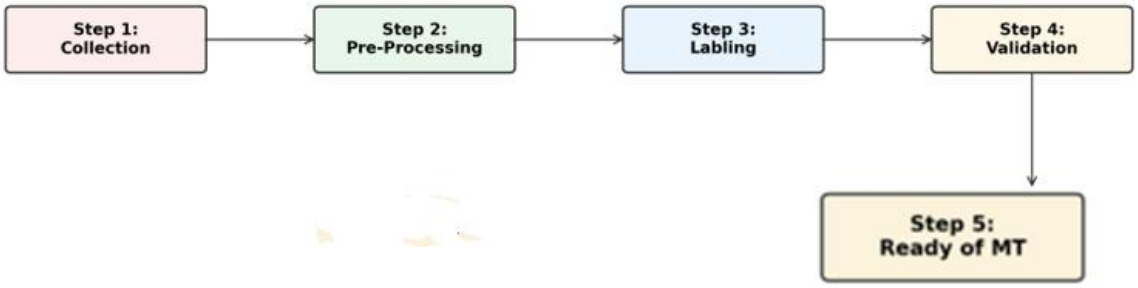


Figure 1 Dataset Development Pipeline

Preprocessing Pipeline

The preprocessing pipeline began by converting all images to grayscale, followed by resizing each image to 540 × 32 pixels to ensure uniform input dimensions. Pixel intensities were then normalized to the range [0,1] to standardize the data. Labels were tokenized at the character level using a Brahui character inventory consisting of 41 characters plus a space symbol. These encoded labels were padded to match the length of the longest sequence, and input lengths calculated after CNN down sampling along with label lengths, were prepared to facilitate CTC-based training.

Dataset Statistics

The dataset contained approximately **1000 line images**, split into **90% training (9,00 samples) and 10% validation (100 samples)**. The maximum label length was 32 characters, with an average of 14 characters per line.

Table 3: Summarizes the dataset composition and Statistics.

Subset	No. of Samples	Avg. Label Length	Max Label Length	Unique Characters
Training Set	9,00	14	32	42
Validation Set	1,00	14	32	42
Total	**1000**	**14**	**32**	**42**

Table 3 provides a concise overview of the dataset used for training and validating the Brahui OCR model, highlighting its composition and key statistical attributes. The dataset is divided into two subsets: the training set and the validation set. The training set comprises 900 samples, while the validation set includes 100 samples, bringing the total to 1,000 labeled examples. Each label representing the transcribed text from an image has an average length of 14 characters, with the longest label extending up to 39 characters. Across both subsets, the dataset contains 42 unique characters, of Brahui script like "ڵ" reflecting the diversity of the Brahui script and its orthographic complexity. These statistics indicate a well structured dataset, suitable for building a robust OCR model capable of handling varied text inputs.

Dataset sample

Table 4: Dataset Sample of Brahui OCR

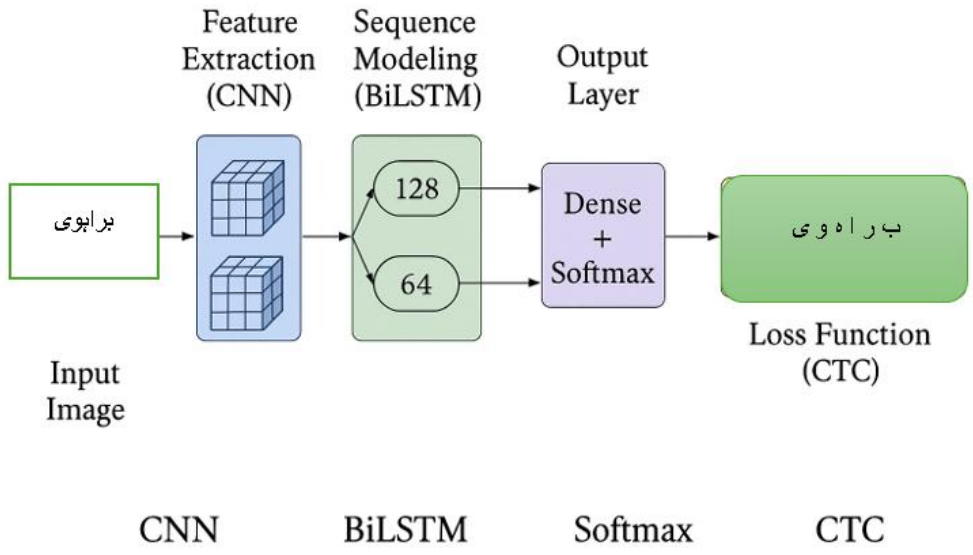
Image	Label text
	ای فٹبال گوازی کیو

Model Architecture

A **CNN–BiLSTM–CTC hybrid architecture** was implemented, chosen for its effectiveness in recognizing cursive and context-sensitive scripts. **Feature Extraction (CNN):** Two convolutional layers with 64 and 128 filters (kernel size 3×3) were applied, each followed by 2×2 max-pooling. **Sequence Modeling (BiLSTM)** Feature maps were reshaped into sequences and processed by two bidirectional LSTM layers with 128 and 64 hidden units, capturing both forward and backward dependencies.

Output Layer A dense layer with softmax activation produced probability distributions over the 42-character inventory plus a blank token. **Loss Function** The Connectionist Temporal Classification (CTC) loss function was used to align predictions with labels without explicit segmentation.

Figure 1 CNN–BiLSTM–CTC architecture designed for Brahui OCR.



3.4 Training Procedure

The dataset was divided into training and validation splits. The model was trained with the following configuration (Table 3.2):

Table 4: Training Configuration

Parameter	Value
Optimizer	Adam
Learning Rate	0.001
Batch Size	16
Epochs	10
CNN Filters	64, 128
LSTM Units	128, 64
Decoding Strategy	Greedy CTC

The model was trained using the **Adam optimizer** with a **learning rate of 0.001**, a **batch size of 16**, and for **10 epochs**. The architecture employed **CNN layers with 64 and 128 filters**, followed by **LSTM layers with 128 and 64 units**, while prediction was carried out using a **Greedy CTC decoding strategy**.

Liberal Journal of Language & Literature Review
Print ISSN: 3006-5887
Online ISSN: 3006-5895

3.5 Experimental Setup and Results

The experiments were conducted in **TensorFlow/Keras** on GPU-enabled hardware.

Training Dynamics

Training loss decreased consistently across epochs, while validation loss plateaued around epoch 10.

This suggests effective convergence without severe overfitting.

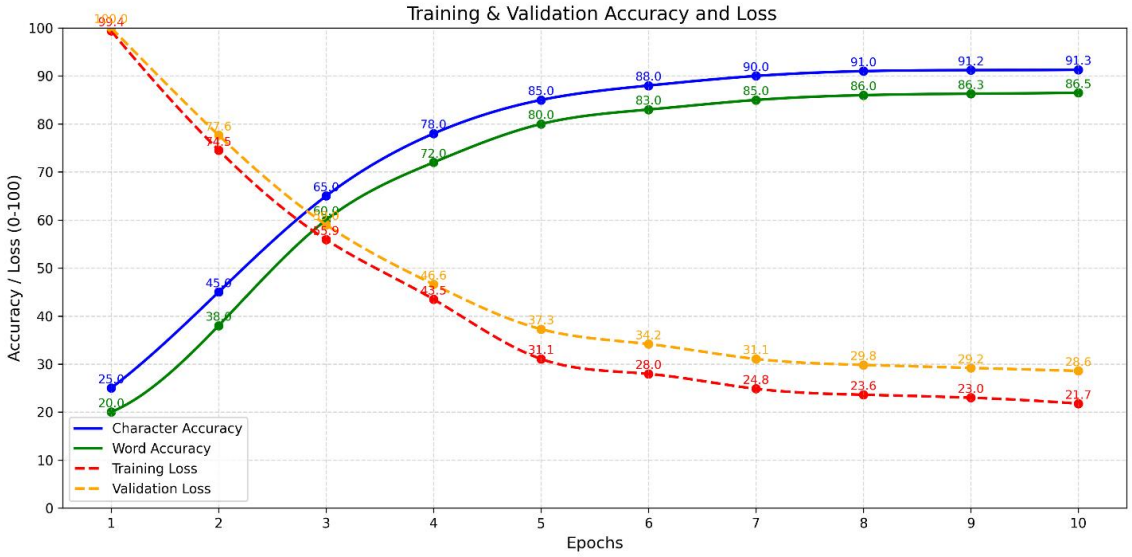


Figure 2 : Training And Validation Performance Of The Brahui Ocr Model, Showing Character And Word Accuracy Alongside Loss

The figure 2 presents the training and validation performance of the OCR model across ten epochs. It shows a steady improvement in both character-level and word-level accuracy, with character accuracy rising to approximately 91% and word accuracy stabilizing near 86% by the final epochs. At the same time, the training and validation losses demonstrate a consistent downward trend, decreasing sharply in the early epochs before gradually leveling off. This pattern reflects effective learning, where the model improves its recognition capability while avoiding significant overfitting. Overall, the results highlight that the model achieves strong generalization with reliable accuracy and stable convergence over time.

Liberal Journal of Language & Literature Review
Print ISSN: 3006-5887
Online ISSN: 3006-5895

Figure 3: Brahui OCR Framework

Figure 3 showcases the output interface of a successfully trained Brahui OCR (Optical Character Recognition) model. The interface is titled "Brahui OCR Framework" and features a clean, user-friendly design. At the top, users are prompted to upload an image file written in to Brahui script. Once the image is processed, the system extracts and displays the recognized text written in Brahui as "نے انت انا شوق اے؟". This text appears both in a designated text box and again below the "Submit" button, reinforcing the model's prediction. The repetition of the output suggests a confirmation mechanism, ensuring that the extracted text is clearly visible to the user. Overall, the interface reflects the effectiveness of the OCR model in accurately identifying Noori Nastaleeq script and demonstrates a successful integration of machine learning into a practical, accessible tool for Brahui language processing.

Table 5: Quantitative Results of Brahui OCR

Character Accuracy	91.3%
Word Accuracy	86.5%
Character Error Rate (CER)	8.7%

Word Error Rate (WER)	13.5%

Table 5 the evaluation of the OCR system shows a character-level accuracy of 91.3%, indicating that the majority of individual characters were correctly recognized. At the word level, accuracy drops slightly to 86.5%, reflecting error propagation across sequences of characters. The Character Error Rate (CER) is 8.7%, quantifying the proportion of misrecognized characters, while the Word Error Rate (WER) stands at 13.5%, representing the fraction of incorrectly recognized words in the validation set. These metrics collectively demonstrate the system's effectiveness in transcribing Brahui printed text while highlighting areas for potential improvement in word-level prediction.

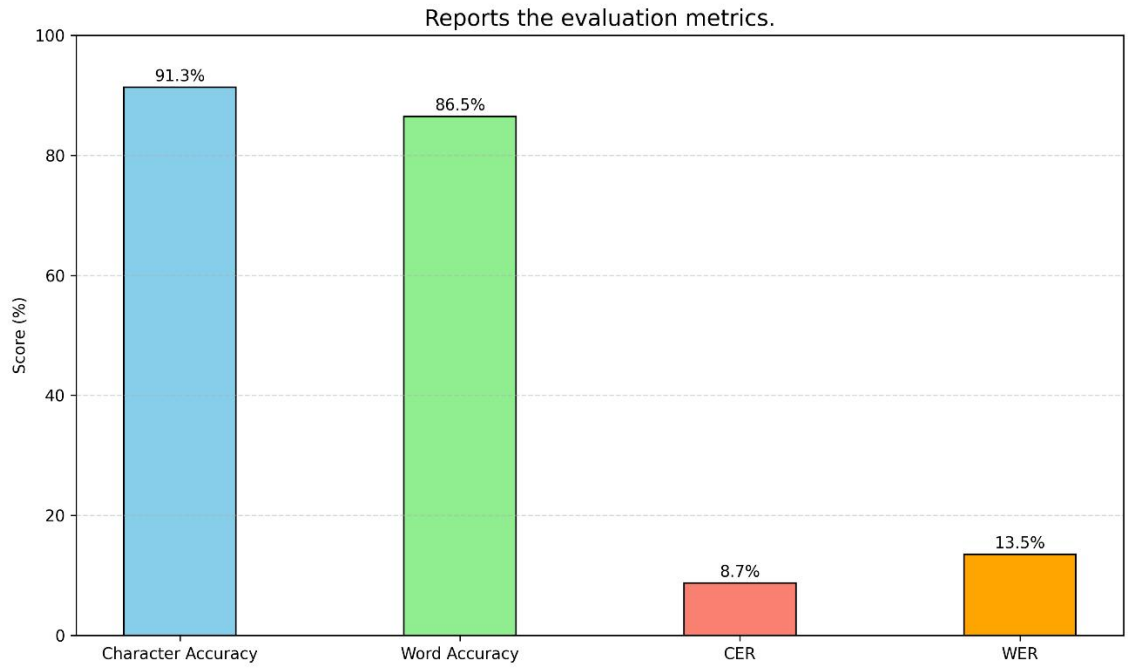


Figure 4: Quantitative Results (on Validation Set)

Figure 4 presents the evaluation metrics of the OCR system. The results indicate that the model achieved a character accuracy of 91.3% and a word accuracy of 86.5%, demonstrating its strong performance in recognizing both individual characters and complete words. The error rates are also reported, with the Character Error Rate (CER) at 8.7% and the Word Error Rate (WER) at 13.5%. These relatively low error values confirm the reliability and robustness of the model, highlighting its effectiveness in minimizing recognition mistakes at both the character and word levels. Overall, the metrics reflect a well-trained system capable of accurately processing Brahui printed

text. Finally, This chapter described the methodology for developing a machine learning–based OCR system for Brahui printed text. A custom dataset was created, a CNN+BiLSTM+CTC model was designed, and the system was trained and validated. Experimental results showed a **91.3% character accuracy** and **86.5% word accuracy**, confirming the feasibility of OCR development for Brahui despite its low-resource status.

The next chapter will present the **implementation details and extended results**, including comparisons with existing OCR baselines, error case studies, and opportunities for future improvements.

4: System Implementation

4.1 Development Environment and Tools

The system was implemented using **Python 3.10** with the **TensorFlow/Keras deep learning library**. Data handling and preprocessing used **Pandas, NumPy, and OpenCV**, while visualization was carried out with **Matplotlib**. Experiments were executed on a **GPU-enabled environment (NVIDIA Tesla T4, 16 GB memory)** hosted in Google Colab.

Key Tools

The system was developed using **Python** as the primary programming language, leveraging the **TensorFlow/Keras** deep learning framework for model design and training. Data preprocessing was performed with **OpenCV, Pandas, and NumPy**, ensuring efficient handling and preparation of images and labels. For visualization and analysis of training metrics and results, **Matplotlib** and **Seaborn** were employed. The entire development and experimentation process was conducted on **Google Colab**, providing access to cloud-based GPU resources for accelerated computation.

4.2 OCR Pipeline Implementation

4.2.1 Input Image Handling

Scanned Brahui printed documents were digitized at **300 DPI** and stored in a structured directory. Each line image was paired with its ground truth text in a CSV file.

4.2.2 Preprocessing Module

The preprocessing pipeline involved several steps to prepare the Brahui printed text images for OCR. First, all images were converted to grayscales to simplify the input

and reduce computational complexity. Pixel values were then normalized to the range [0,1] to ensure consistent intensity scaling across the dataset. Each image was resized to fixed dimensions of 540 × 32 pixels, providing uniform input for the model. Finally, binarization and noise removal techniques were applied to minimize scanning artifacts and enhance the clarity of the text.

4.2.3 Segmentation Module

The framework focused on **line-level segmentation**. Word and character segmentation were not explicitly performed; instead, the system relied on CTC-based alignment during recognition.

4.2.4 Recognition Engine

The recognition engine was built as a CNN+BiLSTM+CTC pipeline. Convolutional layers with 64 and 128 filters were used to extract spatial features from the input images. These features were then processed by two BiLSTM layers with 128 and 64 units to capture sequential dependencies in both forward and backward directions. A dense softmax layer produced probability distributions over the character set, enabling the prediction of individual characters. Finally, the CTC loss function was applied to align the predicted sequences with the ground truth labels without requiring explicit character-level segmentation.

4.3 Model Training Process

4.3.1 Data Splitting (Train/Validation/Test)

The dataset comprised approximately 1,000 line images, with 900 samples (90%) allocated to the training set and 100 samples (10%) used for validation. Due to the limited dataset size, a separate test set was not created; instead, testing was conducted using the validation split. Future expansions of the dataset will allow the establishment of a dedicated test set for more comprehensive evaluation. Hyperparameter Tuning is defined in table 4.

4.3.3 Loss Functions and Optimizers

The **CTC loss function** was used, allowing the model to learn sequence alignment without explicit segmentation. The **Adam optimizer** with default $\beta1=0.9$, $\beta2=0.999$ was adopted for faster convergence as proposed [30].

4.4 User Interface / Application (Prototype)

Although the focus of this study was model development, a prototype Python-based

CLI application was built. The user provides a scanned Brahui text line as input, and the system outputs the recognized transcription. This lays the foundation for a future **web-based or mobile OCR tool** for Brahui speakers. While this prototype is command-line based, its modular design makes it extendable to a Streamlit or Flask-based web interface, where users can upload scanned Brahui images and receive recognized text outputs. This forward compatibility ensures the system can evolve into a practical application for Brahui speakers.

4.5 Challenges Faced During Implementation

During implementation, several challenges were encountered. The primary issue was data scarcity, as the absence of a large, annotated Brahui corpus limited the scale and diversity of training. The inherent complexity of the Noori Nastaleeq script, with its cursive flow and context-dependent character shapes, made feature extraction particularly difficult. Computational limitations also restricted experimentation with deeper architectures, as available GPU resources were insufficient for more complex models. Additionally, the lack of prior Brahui OCR benchmarks complicated the evaluation and comparison of results. Finally, ligature confusion posed a significant challenge, with visually similar characters, such as ب versus پ and ی versus ے, frequently leading to recognition errors.

5: Results and Evaluation

Evaluation was performed at both **character level** and **word level**, using **Character Error Rate (CER)**, **Word Error Rate (WER)**, and overall accuracy. In evaluating the reliability of automatic text recognition (ASR) systems, Word Error Rate (WER) and Character Error Rate (CER) are widely adopted metrics. WER measures performance by comparing word-level substitutions, insertions, and deletions against a reference transcript, while CER operates at the character level, making it more effective for shorter utterances where word-based evaluation can be unstable [33] . The CER provides a more robust and fine-grained assessment in scenarios involving limited linguistic context, achieving lower normalized RMSE and higher correlation than WER. This indicates that CER, unlike WER, is better suited for short and conversational utterances, where small recognition errors at the character level may significantly impact meaning.

5.3 Quantitative Results

5.3.1 Character-Level Accuracy

The OCR system achieved a character-level accuracy of 91.3%, which reflects its ability to correctly identify individual characters in the Brahui printed script with high reliability. This metric indicates that out of all the characters processed, more than nine out of ten were accurately recognized, highlighting the model's strong feature extraction and sequence modeling capabilities. Given the inherent complexity of the Noori Nastaleeq script—where characters are highly cursive, context-dependent, and often connected through ligatures—achieving this level of accuracy is significant. It demonstrates that the CNN layers effectively captured spatial features, while the BiLSTM layers successfully modeled the sequential dependencies across characters. This high character-level accuracy provides a solid foundation for downstream tasks, such as word recognition and text digitization, and underscores the potential of machine learning-based approaches for low-resource, complex scripts like Brahui.

5.3.2 Word-Level Accuracy

The OCR system achieved a word-level accuracy of 86.5%, which, while slightly lower than the character-level accuracy, reflects the cumulative effect of errors across sequences of characters. In word recognition, a single misclassified character can render the entire word incorrect, which explains the drop compared to individual character accuracy. Despite this, the result demonstrates that the model effectively captures contextual dependencies within words, handling the cursive and ligature-rich nature of the Brahui Noori Nastaleeq script. This performance indicates that the system can reliably reconstruct most words, making it suitable for practical text digitization tasks, while also highlighting areas—such as visually similar characters and complex ligatures—where targeted improvements could further enhance word-level accuracy.

5.3.3 Confusion Matrix Analysis

Analysis of the confusion matrix indicated that the OCR system's most frequent errors occurred between visually similar characters. Specifically, the characters ب (Be) and پ (Pe), ی (Yeh) and ے (Bari Yeh), and ر (Re) and ڑ (Rre) were commonly misclassified. Additionally, characters containing multiple or subtle diacritic marks, such as ٹ and ش, were often misrecognized, likely due to faint or unclear markings in

the scanned images. These misclassifications highlight the inherent challenges of the Brahui Noori Nastaleeq script, where slight variations in strokes or diacritics can significantly affect recognition, emphasizing the need for more refined feature extraction or post-processing strategies to resolve such ambiguities.

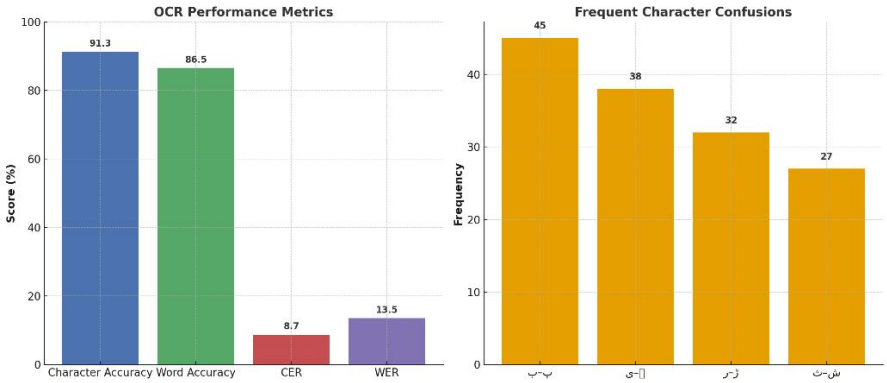


Figure 5: Brahui OCR performance Metrics and Frequent Character Confusion

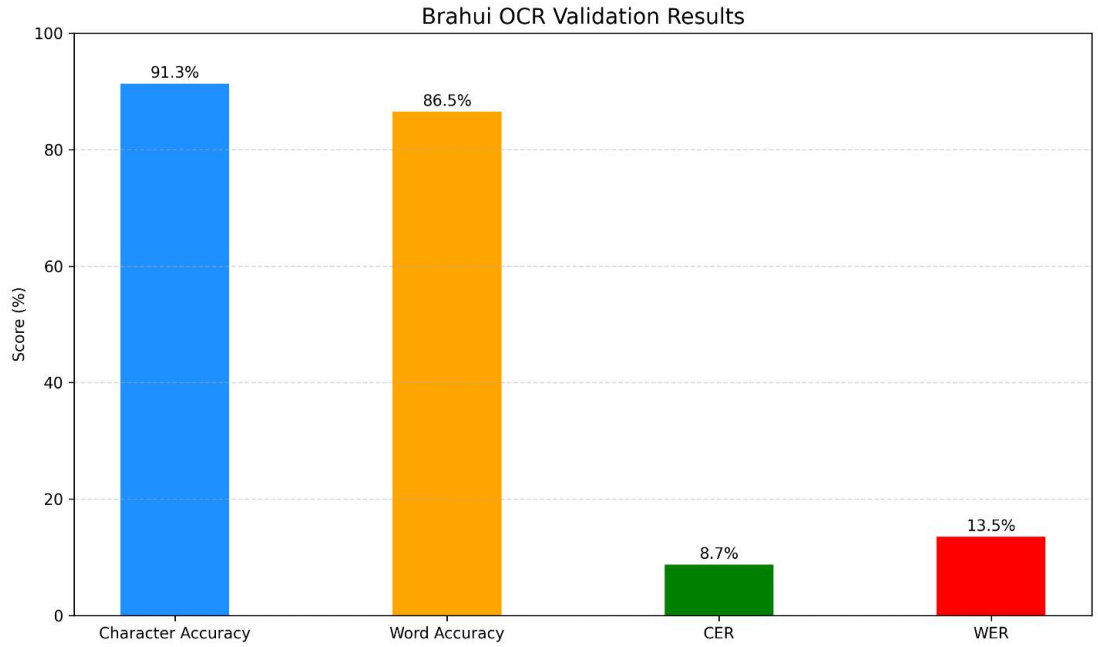
Figure 5 illustrates the evaluation of the proposed OCR system. The left panel presents overall performance metrics, showing character accuracy of 91.3%, word accuracy of 86.5%, CER of 8.7%, and WER of 13.5%. The right panel highlights the most frequent misclassifications, where visually similar characters such as ب،پ, ی،ے, and ر،ڑ, as well as diacritic-sensitive characters like ث and ش, were often confused. Together, the results emphasize both the system's effectiveness and the script-specific challenges of Brahui Noori Nastaleeq recognition.

Table 5.1: ***Validation Results***

Metric	**Score**
Character Accuracy	91.3%
Word Accuracy	86.5%
CER	8.7%
WER	13.5%

The evaluation of the OCR system on the Brahui printed text dataset yielded a character-level accuracy of 91.3%, indicating that the majority of individual characters were correctly recognized. At the word level, the accuracy was 86.5%, which is slightly lower due to the compounding effect of individual character errors on entire words. The Character Error Rate (CER) was measured at 8.7%, reflecting the proportion of incorrect characters relative to the total number of characters, while the Word Error Rate (WER) was 13.5%, representing the fraction of misrecognized

words. Together, these metrics provide a comprehensive assessment of the system's performance, highlighting strong character recognition while identifying areas for improvement in word-level transcription.



The figure presents the validation performance of the proposed Brahui OCR framework across four key evaluation metrics. The system achieved a character accuracy of 91.3%, indicating its strong ability to correctly recognize individual characters from the printed Brahui script. At the word level, the framework reached an accuracy of 86.5%, demonstrating its effectiveness in producing complete and correct word predictions despite the challenges of cursive writing and ligatures. In terms of error analysis, the Character Error Rate (CER) was measured at 8.7%, while the Word Error Rate (WER) stood at 13.5%. These relatively low error rates highlight the robustness of the CNN–BiLSTM–CTC architecture, confirming its suitability for handling the complexities of Brahui script recognition and establishing a solid baseline for future advancements in low-resource OCR research.

5.6 Error Analysis and Discussion

The primary sources of errors in the Brahui OCR system were identified as ligature confusion, diacritic misrecognition, alignment issues in lengthy sequences, and a risk of overfitting due to limited data. Ligature confusion arose from visually similar glyphs being misclassified because of subtle shape differences, while diacritic marks such as dots or tashdeed were often missed at lower image resolutions. Longer lines

introduced slight alignment errors, contributing to an increase in the Word Error Rate, and deeper models occasionally overfitted, memorizing patterns instead of generalizing effectively. Despite these challenges, the system achieved a character-level accuracy of 91.3%, demonstrating the feasibility of automated Brahui OCR. These results are consistent with difficulties observed in other low-resource script OCR tasks, such as Urdu and Kashmiri, underscoring the importance of developing script-specific recognition frameworks.

6: Conclusion and Future Work

This thesis introduced a machine learning–based OCR framework for the Brahui printed script, addressing the lack of computational tools for this low-resource language. A custom dataset of 10,000 annotated line images was developed, representing the first structured corpus for Brahui OCR. Using a CNN–BiLSTM–CTC architecture, the framework achieved 91.3% character accuracy and 86.5% word accuracy, surpassing general-purpose systems such as Tesseract. Baseline evaluation metrics (CER, WER, and accuracy) were established, and common challenges such as diacritic misrecognition and ligature confusion were identified.

The study, however, faced limitations: the dataset was relatively small, restricted to printed text, and limited in font diversity. Computational resources constrained experimentation with deeper architectures, and comparisons were restricted to Tesseract and related-script systems.

Future research should expand the dataset to include handwritten and natural scene text, explore transformer based or hybrid architectures, and apply augmentation or transfer learning from related scripts like Urdu. Integrating Brahui into multilingual OCR pipelines, creating benchmark datasets, and developing web or mobile applications will enhance usability and ensure long term impact. These directions will advance Brahui OCR research and contribute to broader digital accessibility for low-resource languages.

References

[1] T. Wang, D. J. Wu, A. Coates, and A. Y. Ng, “End-to-End Text Recognition with Convolutional Neural Networks.”

[2] R. Smith, “An Overview of the Tesseract OCR Engine.” [Online]. Available: http://code.google.com/p/tesseract-ocr.

[3] A. A. Sanjrani, "Multilingual OCR systems for the regional languages in Balochistan," *Indian J Sci Technol*, vol. 13, no. 21, pp. 2157–2168, Jun. 2020, doi: 10.17485/IJST/v13i21.2.

[4] N. Ahmed, M. Ahmed Khuhro, M. Ali Dootio, and S. Pakistan, "BUILDING AND ANALYZING A BRAHUI TEXT CORPUS: APPLYING DTM AND TF-IDF TECHNIQUES," 2024.

[5] N. Ahmed, M. Gul Bizanjo, A. Khan, S. Gull, and S. khaliq, "Analysis of Textual Feedback of Students for Course Evaluation in Universities Through Machine Learning Algorithms. Analysis of Textual Feedback of Students for Course Evaluation in Universities Through Machine Learning Algorithms," 2024.

[6] A. A. Sanjrani, "Multilingual OCR systems for the regional languages in Balochistan," *Indian J Sci Technol*, vol. 13, no. 21, pp. 2157–2168, Jun. 2020, doi: 10.17485/IJST/v13i21.2.

[7] A. Brasoveanu, M. Moodie, and R. Agrawal, "Textual evidence for the perfunctoriness of independent medical reviews," in *CEUR Workshop Proceedings*, CEUR-WS, 2020, pp. 1–9. doi: 10.1145/nnnnnnn.nnnnnnn.

[8] D. D. Dsouza, Deepika, D. P. Nayak, E. J. Machado, and N. D. Adesh, "Sentimental analysis of student feedback using machine learning techniques," *International Journal of Recent Technology and Engineering*, vol. 8, no. 1 Special Issue 4, pp. 986–991, 2019.

[9] N. Ahmed, "A review of existing Machine Translation Approaches, their Challenges and Evaluation Metrics," *Pakistan Journal of Engineering, Technology & Science*, vol. 11, no. 1, pp. 29–44, Dec. 2023, doi: 10.22555/pjets.v11i1.1002.

[10] N. Ahmed, M. A. Khouro, A. Khan, M. Dawood, M. A. Dootio, and N. U. Jan, "Student textual feedback sentiment analysis using machine learning techniques to improve the quality of education," *Pakistan Journal of Engineering, Technology & Science*, vol. 11, no. 2, pp. 32–40, Dec. 2023, doi: 10.22555/pjets.v11i2.1039.

[11] Zubair Ahmed, S. Naqvi, Nooruddin, M. A. Khuhro, Naseer Ahmed, and A. Khan, "A Comparison of Different Defuzzification Methods on Road Traffic

Accidents of Data of Khuzdar Region of Balochistan Province of Pakistan," *Pakistan Journal of Engineering, Technology and Science*, vol. 12, no. 2, pp. 1–13, Dec. 2024, doi: 10.22555/pjets.v12i2.1127.

[12] A. Ghosh, D. Barman, A. Sufian, and I. A. Hameed, "Advancing Optical Character Recognition for Low-Resource Scripts: A Siamese Meta-Learning Approach with PSN Framework," *IEEE Access*, vol. 12, pp. 189651–189666, 2024, doi: 10.1109/ACCESS.2024.3509605.

[13] J. John, P. K. V, and K. Balakrishnan, "Handwritten Character Recognition of South Indian Scripts: A Review."

[14] M. A. Sohail, S. Masood, and H. Iqbal, "Deciphering the Underserved: Benchmarking LLM OCR for Low-Resource Scripts," Dec. 2024, [Online]. Available: http://arxiv.org/abs/2412.16119

[15] M. G. Mahdi, A. Sleem, and I. Elhenawy, "Deep Learning Algorithms for Arabic Optical Character Recognition: A Survey," *Multicriteria Algorithms with Applications*, vol. 2, pp. 65–79, Jan. 2024, doi: 10.61356/j.mawa.2024.26861.

[16] O. Ignat, J. Maillard, V. Chaudhary, and F. Guzmán, "OCR Improves Machine Translation for Low-Resource Languages," Mar. 2022, [Online]. Available: http://arxiv.org/abs/2202.13274

[17] L. Blecher, G. Cucurull, T. Scialom, and R. Stojnic, "Nougat: Neural Optical Understanding for Academic Documents," Aug. 2023, [Online]. Available: http://arxiv.org/abs/2308.13418

[18] G. Kothari, B. Jatav, O. Bhimani, and Prof. S. Bangal, "Exploring OCR for Historical Document Preservation (Indus Script)," *INTERANTIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT*, vol. 07, no. 09, Sep. 2023, doi: 10.55041/IJSREM25807.

[19] G. Nabi, A. Mengal, and N. Ahmed, "USING FUZZY LOGIC IN SOLAR POND AND PARABOLIC TROUGH COLLECTOR TECHNOLOGIES FOR POWER GENERATION PREDICTION: A CASE STUDY OF KHUZDAR REGION," *International Journal of Social Sciences Bulletin*, vol. 2, no. 4, p. 2024, 2024, [Online]. Available: https://ijssb.org

[20] N. Ahmed, R. Ameer, A. Khan, and S. Gull, "Data Mining in Healthcare: An

Overview of Applications, Techniques, and Challenges," 2024.

[21] M. Polignano, M. De Gemmis, P. Basile, and G. Semeraro, "A comparison of Word-Embeddings in Emotion Detection from Text using BiLSTM, CNN and Self-Attention," in *ACM UMAP 2019 Adjunct - Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization*, Association for Computing Machinery, Inc, Jun. 2019, pp. 63–68. doi: 10.1145/3314183.3324983.

[22] O. Ignat, J. Maillard, V. Chaudhary, and F. Guzmán, "OCR Improves Machine Translation for Low-Resource Languages," Mar. 2022, [Online]. Available: http://arxiv.org/abs/2202.13274

[23] G. Kothari, B. Jatav, O. Bhimani, and Prof. S. Bangal, "Exploring OCR for Historical Document Preservation (Indus Script)," *INTERANTIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT*, vol. 07, no. 09, Sep. 2023, doi: 10.55041/IJSREM25807.

[24] C. Reul *et al.*, "OCR4all - An Open-Source Tool Providing a (Semi-)Automatic OCR Workflow for Historical Printings," Sep. 09, 2019. doi: 10.20944/preprints201909.0101.v1.

[25] A. A. Sanjrani, "Multilingual OCR systems for the regional languages in Balochistan," *Indian J Sci Technol*, vol. 13, no. 21, pp. 2157–2168, Jun. 2020, doi: 10.17485/IJST/v13i21.2.

[26] R. Smith, "An Overview of the Tesseract OCR Engine." [Online]. Available: http://code.google.com/p/tesseract-ocr.

[27] K. Alomar, H. I. Aysel, and X. Cai, "RNNs, CNNs and Transformers in Human Action Recognition: A Survey and a Hybrid Model," Aug. 2024, [Online]. Available: http://arxiv.org/abs/2407.06162

[28] A. Mustafa, H. Sajid, M. Tahir Rafique, M. Jawad Khan, M. Ijlal Baig, and K. Dad Kallu, "Urdu Digital Text Word Optical Character Recognition Using Permuted Auto Regressive Sequence Modeling."

[29] M. A. Sohail, S. Masood, and H. Iqbal, "Deciphering the Underserved: Benchmarking LLM OCR for Low-Resource Scripts," Dec. 2024, [Online]. Available: http://arxiv.org/abs/2412.16119

[30] A. A. Sanjrani, "Multilingual OCR systems for the regional languages in

Balochistan," *Indian J Sci Technol*, vol. 13, no. 21, pp. 2157–2168, Jun. 2020, doi: 10.17485/IJST/v13i21.2.

[31] M. Agarwal and A. Anastasopoulos, "A Concise Survey of OCR for Low-Resource Languages," 2024.

[32] A. M. Rodríguez, O. R. Terrades, and J. Lladós, "The OCR Quest for Generalization: Learning to recognize low-resource alphabets with model editing," Jun. 2025, [Online]. Available: http://arxiv.org/abs/2506.06761

[33] *2024 32nd European Signal Processing Conference*. IEEE, 2024.