Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X

Volume 4, Issue 3, 2026

A SMART OCR FRAMEWORK FOR DIGITIZATION OF URDU-BASED DOCUMENTS OF DISTRICT EDUCATION AUTHORITIES IN PUNJAB, PAKISTAN

Muhammad Naeem[1], Muhammad Umer Hayat[2], Khalid Hussain[3], Ahmad Khan[4]

*[*1,2,3,4]Faculty of Computer Science and Information Technology, The Superior University Lahore, Pakistan*

[*1]lastleaf36@gmail.com, [2]look2umer@yahoo.com, [3]khalidhussain.fsd@superior.edu.pk, [4]ahmad.khan.fsd@superior.edu.pk

DOI: https://doi.org/10.5281/zenodo.19233218

Keywords
Urdu OCR; Nastaliq script; District Education Authorities; educational administration; document digitization; intelligent document processing; low-resource language computing.

Article History
Received: 26 January 2026
Accepted: 09 March 2026
Published: 26 March 2026

Corresponding Author: *
Muhammad Naeem

Abstract

The digitization of administrative records is essential for efficient governance, transparency and data-driven decision-making. In Punjab, Pakistan, District Education Authorities (DEAs) are responsible for managing public-sector schools across thirty-six districts and generate a large volume of official correspondence, including circulars, notifications, directives and policy-related letters. A substantial portion of this communication is produced in Urdu to ensure accessibility for Class-IV employees, local stakeholders and School Management Council members who may have limited proficiency in English. However, most of these Urdu documents remain stored as paper files or scanned images, which restricts their searchability, preservation and analytical reuse.

The digitization of such records is technically challenging because Urdu is commonly written in the Nastaliq script, which is cursive, context-sensitive and visually complex. Character shape variation, ligature formation, diacritics, variable baselines, degraded scans, inconsistent layouts, stamps, signatures and handwritten annotations significantly reduce the effectiveness of conventional OCR systems. The problem is further intensified by the presence of formal, legal, procedural and domain-specific educational vocabulary that is not adequately represented in general-purpose OCR datasets.

This paper addresses these challenges by proposing a smart OCR framework tailored to Urdu-based letters issued by District Education Authorities in Punjab. The proposed framework is intended to support the recognition of Urdu circulars, notifications and official letters while facilitating searchable archiving, improved institutional recordkeeping and metadata-oriented document management. By focusing on a low-resource script and a high-value administrative domain, the study contributes to both Urdu language technology and the broader digital transformation of educational governance in Pakistan.

1. INTRODUCTION

The growing use of information and communication technologies has reshaped how institutions manage, preserve and retrieve official records. Digitization is now central to efficient administration, organizational transparency and evidence-based planning [1]. Nevertheless, in many developing settings, including Pakistan, a substantial amount of official documentation still exists in paper-based or image-based form, particularly in regional and national languages. Such dependence on manual or semi-digital records limits accessibility, slows retrieval, weakens long-term preservation and constrains large-scale information analysis. Within this context, Optical Character Recognition (OCR) has emerged as an enabling technology for transforming scanned documents into machine-readable and searchable text [2].

This challenge is especially important in the public education system of Punjab, where district-level administration is carried out through thirty-six District Education Authorities (DEAs). These authorities oversee government schools and routinely issue official documents such as letters, notifications, circulars, directives, appointment orders, transfer orders, disciplinary notices and financial instructions. Because the educational workforce and associated stakeholders come from diverse social and educational backgrounds, Urdu functions as a practical and inclusive medium of official communication. It is particularly important for Class-IV employees and School Management Council (SMC) members, many of whom may have limited English proficiency. As a result, Urdu-based administrative correspondence has become a significant part of educational governance in Punjab.

Over time, DEAs have accumulated large collections of Urdu documents containing administrative, legal, procedural and technical information relevant to policy implementation, staff management, financial oversight, audits and institutional continuity. Despite their importance, most of these records remain confined to physical files or unstructured scanned copies. This creates serious operational limitations, including slow retrieval, weak archival organization, restricted reuse and vulnerability to physical deterioration. In an era increasingly shaped by digital governance and data-driven administration, the continued dependence on non-searchable Urdu records restricts institutional efficiency and limits the value of historical documentation [3][4].

Although OCR offers a promising route toward digitization, Urdu remains a difficult script for automated recognition [3]. Urdu is commonly written in Nastaliq, a highly cursive script characterized by context-sensitive letter forms, complex ligatures, slanted text flow and variable baselines [5]. In addition, diacritics and dots carry semantic significance and require accurate visual recognition. Due to degraded scanning quality, variable layout of the documents, stamps, signatures and handwritten notes on official documents this task becomes more difficult [2]. Existing OCR systems, especially those developed for Latin-script or general-purpose applications, are not sufficiently robust for such conditions [4]. Even Urdu-oriented systems often struggle with specialized educational and administrative vocabulary, reducing their reliability in real-world institutional settings [6].

The central research problem, therefore, lies in the absence of a dependable OCR solution tailored to Urdu-based educational administrative documents. There is a clear need for a domain-specific framework capable of handling the linguistic, visual and structural complexities of official Urdu correspondence while also supporting archival access and intelligent document management. In response, this study proposes a smart OCR framework for Urdu-based letters issued by District Education Authorities in Punjab, Pakistan. Developed for educational administration, this framework focuses on precise OCR, efficient digitization of the documents with searchable archiving facility and an organized document management system. By situating OCR research within the operational realities of Punjab's public education system, the study contributes to low-resource language computing as well as to the modernization of public-sector record management.

2. Literature Review

The advanced deep learning frameworks have changed the earlier rule-based and template-matching methods which were used in OCR [5]. Early OCR systems relied on handcrafted features and performed effectively on clean, printed Latin text but struggled with cursive and complex scripts [6]. The development of modern OCR systems, including the Tesseract engine [5] and connectionist sequence models, enabled improved recognition of unconstrained text [2]. Deep learning models such as CNNs, RNNs and transformers—now drive the OCR, improve the feature extraction, sequence modeling and increasing the accuracy on complex scripts. Transformer-based OCR models, in particular, demonstrate strong performance in capturing contextual dependencies in cursive text [7]. Evaluation metrics such as Character Error Rate (CER) and Word Error Rate (WER) are commonly used to assess OCR performance in low-resource settings [6].

OCR for Arabic-script languages introduces additional challenges due to cursive writing, contextual character variations and diacritics [8]. Studies on Arabic OCR highlight issues related to segmentation ambiguity, ligature complexity and degraded image quality [9], [10]. These challenges are directly relevant to Urdu, which shares the same script base but is typically written in the more complex Nastaliq style [3].

Urdu Nastaliq script presents unique difficulties due to its diagonal writing style, overlapping ligatures and variable baselines. Small variations in diacritics and dots can significantly alter meaning, increasing recognition complexity [4]. Traditional OCR approaches have proven inadequate for handling these features, leading to the adoption of machine learning and deep learning methods [11].

Recent research in Urdu OCR has focused on developing specialized models and datasets. Rahman et al. proposed the UTRNet model for the recognition of Urdu text with high resolution, demonstrating the improved accuracy by using the deep neural networks [12]. Similarly, dataset-oriented contributions such as UTRSet highlight the importance of annotated corpora for OCR training and evaluation [12][13]. Survey studies indicate that, despite progress, Urdu OCR remains challenging, particularly for handwritten text and noisy documents [9], [10].

Preprocessing plays a critical role in OCR performance. Techniques such as noise reduction, binarization, skew correction and layout normalization significantly improve recognition accuracy, especially in scanned documents containing stamps and annotations [11]. Additionally, post-processing techniques, including language modeling and dictionary-based correction, help refine OCR outputs by resolving ambiguities [9].

Despite these advancements, a significant gap exists in domain-specific OCR applications. Most research focuses on generic datasets and does not address the unique characteristics of administrative documents. Urdu official documents contain formal language, domain-specific terminology and complex layouts, which are not adequately handled by existing OCR systems.

This gap is particularly evident in the education sector, where District Education Authorities generate large volumes of Urdu documents that remain undigitized. The lack of a specialized OCR framework tailored to such documents highlights the need for domain-specific solutions [13]. This study addresses this gap by proposing an OCR framework designed specifically for Urdu-based educational administrative documents.

3. Materials and Methods

3.1 Research Design

This study employed a systematic structured engineering research design to create and test a smart Optical Character Recognition (OCR) system for administrative and educational documents which are written in Urdu. The methodological design was practical and experiment-driven, combining document image processing, deep learning-based text recognition and context-aware post-processing for the specific challenges of Urdu Nastaliq script used in District Education Authority (DEA) correspondence. The overall framework emphasized reproducibility, robustness and practical deployment in public-sector educational record management [14].

An end-to-end OCR pipeline was built, which covers the phases of data collection, preprocessing, segmentation, recognition, post-processing, metadata extraction and searchable archive of the data. This design was motivated by the limitations of generic OCR systems in handling cursive Urdu text, complex ligatures, diacritics, variable layouts and domain-specific administrative vocabulary [9], [15], [16], [17], [18].

3.2 Problem Formulation and System Objectives

The methodological problem addressed in this study was the absence of a reliable, high-accuracy OCR solution for Urdu administrative documents. Existing OCR systems are either optimized for Latin scripts or trained on limited

Urdu datasets and therefore do not perform adequately on official Urdu documents containing Nastaliq ligatures, technical vocabulary, stamps, signatures and variable formatting [18], [19]. This limitation hinders the digitization, indexing and retrieval of records within public-sector and educational institutions. The system was therefore designed to recognize Urdu circulars, notifications and official letters, while also supporting digital archiving and metadata generation for institutional use.

3.3 End-to-End OCR Pipeline

The proposed OCR pipeline followed a hybrid strategy that integrated traditional image processing with deep learning-based recognition. Document images were first collected from real-world administrative and educational sources to capture variation in page design, font style, print quality and scanning conditions. These images were then subjected to preprocessing operations including noise reduction, binarization, skew correction and layout normalization. After preprocessing, both segmentation-aware and segmentation-free recognition strategies were considered to handle the connected and cursive nature of Urdu Nastaliq script. At the recognition stage, Convolutional Neural Networks (CNNs) were used for visual feature extraction, while sequence modeling was carried out using Long Short-Term Memory (LSTM), Bi-LSTM, or transformer-based architectures. Finally, dictionary-based, rule-based and neural language model techniques were applied for post-processing to improve lexical and contextual correctness.

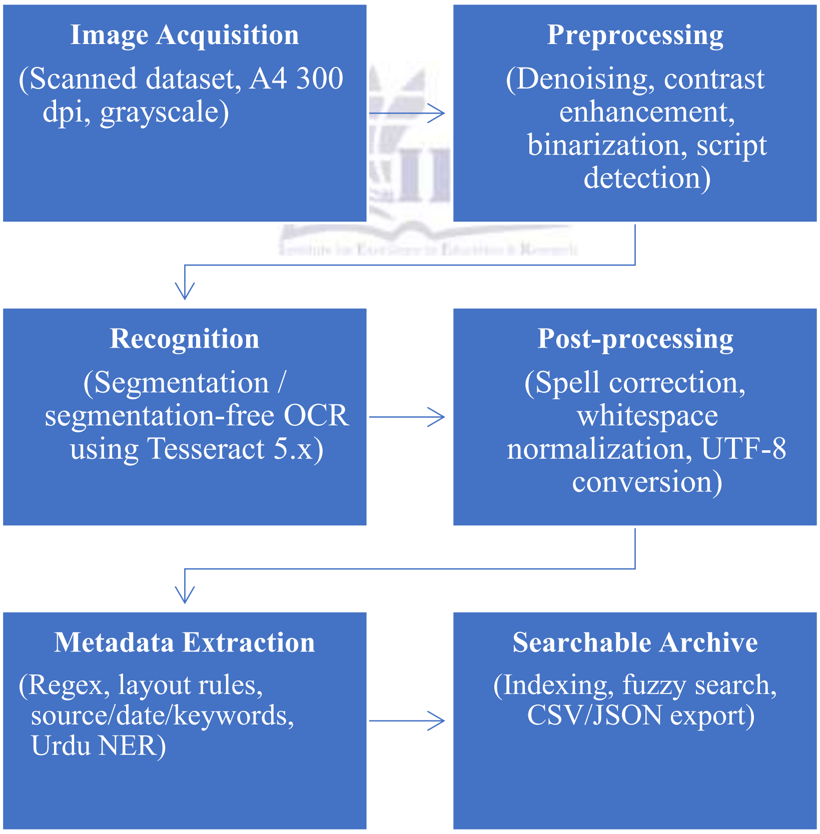
Image Acquisition
(Scanned dataset, A4 300 dpi, grayscale)
Preprocessing
(Denoising, contrast enhancement, binarization, script detection)
Recognition
(Segmentation / segmentation-free OCR using Tesseract 5.x)
Post-processing
(Spell correction, whitespace normalization, UTF-8 conversion)
Metadata Extraction
(Regex, layout rules, source/date/keywords, Urdu NER)
Searchable Archive
(Indexing, fuzzy search, CSV/JSON export)

Figure 1: End-to-end OCR pipeline for Urdu administrative documents

3.4 Data Collection and Dataset Preparation

3.4.1 Data Sources and Collection Strategy

The dataset was prepared to reflect the real-world diversity of Urdu administrative documents. Document images were collected from administrative and educational sources so that the corpus could represent differences in layout, font style, document quality and textual structure. The study intentionally focused on document variability because OCR performance depends strongly on the representativeness, diversity and quality of the dataset. This approach was intended to ensure that the training and testing data resembled actual Urdu official paperwork used in educational administration.

3.4.2 Data Digitization, Annotation and Cleaning

After collection, the documents were digitized and prepared for OCR experiments through structured annotation and cleaning. The dataset preparation pipeline included document organization, text normalization and sample refinement to reduce inconsistencies before model training. Because Urdu administrative documents may contain noise, skew, non-uniform spacing, stamps and mixed formatting, dataset cleaning was treated as a necessary part of model readiness rather than a secondary step.

3.4.3 Dataset Splitting and Augmentation

To enable a systematic system training, its validation and then its testing, the dataset splitting and augmentation were used. Dataset splitting and augmentation were used as part of the preparation framework [21]. The dataset was structured to support reliable model evaluation and better generalization on different conditions of different types of documents.

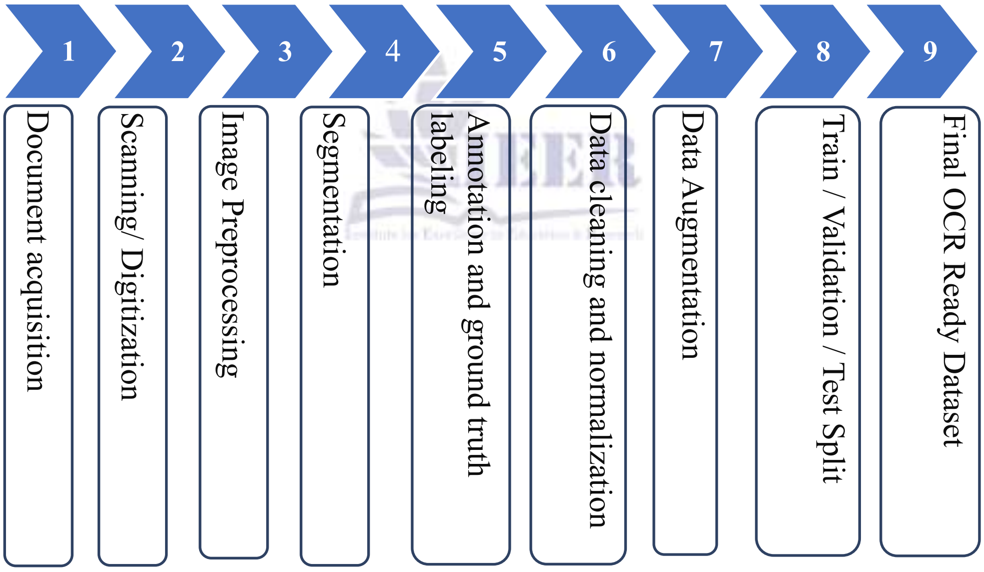

Figure 2: Dataset preparation workflow for Urdu OCR

3.5 Preprocessing Techniques

Preprocessing was a central stage in the proposed OCR system because the quality of scanned administrative documents directly affects downstream recognition performance [13]. The methodology included noise reduction, image enhancement, binarization, skew correction and layout normalization [16]. These operations were applied to improve text visibility, reduce distortions and standardize document appearance before recognition [13], [16]. This was particularly important for Urdu documents because degraded scans, stamps, handwritten notes and variable layouts can increase segmentation and recognition errors.

The preprocessing stage can be summarized as follows:

1) Noise Reduction: removal of scanning artifacts and background disturbances.

2) Binarization: conversion of grayscale or color document images into enhanced foreground–background form.

3) Skew Correction: alignment of slanted or tilted document images.

4) Layout Normalization: standardization of document structure for consistent model input.

```
Input: Raw Document image I
Output: Normalized Image segment S
1.    Convert I to grayscale
2.    Apply Gaussian smoothing
3.    Perform adaptive binarization
4.    Detect and correct skew
5.    Apply morphological operations
6.    Segment text into lines
7.    Normalize and resize segments
8.    Normalize pixel values
      return S
```

Figure 3: Algorithm of Pre-processing

3.6 Segmentation Strategy

The methodology considered both segmentation-aware and segmentation-free approaches. This choice was important because Urdu Nastaliq script is highly cursive and often resists clean separation at the individual character level [22], [23], [24]. Segmenting connected Urdu text into smaller units may introduce recognition errors, especially where ligatures and diacritics are closely packed [25]. Therefore, the system design allowed for explicit line/word-level segmentation where appropriate, while also considering segmentation-free recognition methods capable of processing entire words or lines directly [26].

Table 1
Segmentation approaches used in the developed Urdu OCR system

Aspect	**Explicit Segmentation**	**Segmentation-Free Approach**
Segmentation-aware	Explicit segmentation of lines/words/ligatures before recognition	Useful when layout boundaries are clear
Segmentation- free	Recognition of complete text sequences without explicit unit segmentation	Better suited for connected and cursive Nastaliq text

3.7 OCR Recognition Model

The recognition stage combined visual feature learning with sequence modeling. CNN-based architectures were employed for extracting spatial features from document images, while LSTM, Bi-LSTM and transformer-based models were used for sequence learning and text decoding [20], [27], [28]. This hybrid model design was selected because Urdu OCR requires both robust image-level representation and context-sensitive sequence interpretation. CNNs are effective in learning local visual structures, whereas recurrent and transformer models are better suited to handling dependencies across connected text sequences [29], [30], [31].

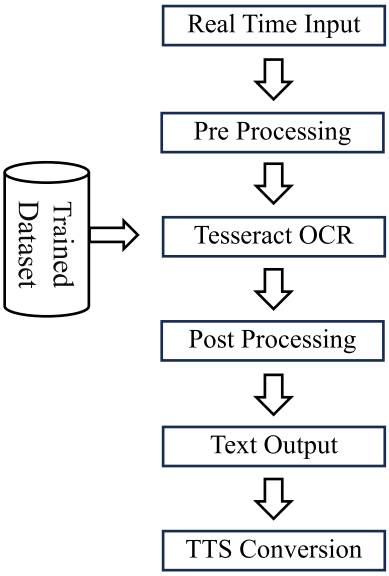

Figure 4: Architecture of the Developed Urdu OCR system

Table 2
Recognition architectures considered in the proposed system

Model Family	Role in OCR pipeline	Expected advantage
CNN	Feature extraction	Captures spatial and structural visual patterns
LSTM / Bi-LSTM	Sequence modeling	Learns contextual dependencies in connected text
Transformer	Attention-based recognition	Handles long-range dependencies and complex sequences

3.8 Text Tokenization and Sequence Representation

Tokenization is important because OCR outputs must be transformed into a machine-processable sequence format for training and decoding. In Urdu OCR, appropriate sequence representation is necessary for handling ligatures, connected writing and word-level dependencies [2], [3], [8], [27].

Table 3
Text tokenization in the Urdu OCR workflow

Component	Description
Image Tensor	H × W × 1 normalized image
Feature Tokens	CNN-extracted embeddings
Tokenized Text	Urdu Unicode characters
Attention Mask	Padding-aware sequence mask
Positional Encoding	Spatial position embeddings

3.9 Post-Processing and Language Modeling

To improve OCR output quality, the system incorporated dictionary-based, rule-based and neural language model-based post-processing. This stage was necessary because even strong visual models may produce orthographically plausible but semantically incorrect outputs, especially in Urdu administrative documents containing technical, legal and formal vocabulary [25], [27]. Context-aware correction was therefore included to improve lexical reliability and preserve document meaning.

The post-processing stage focused on:

- correction of word-level recognition errors,
- restoration of context where ligatures or diacritics were misread,
- and refinement of outputs for domain-specific terminology.

3.10 Integration with Document Management System

A practical contribution of the proposed methodology was the integration of OCR outputs with a Document Management System (DMS). After recognition and correction, the extracted text was transformed into machine-readable and searchable form, allowing indexing, archival access and structured retrieval. This makes the framework more suitable for real-world institutional deployment rather than remaining only a recognition experiment [17], [32].

```
Input: OCR output text $T_i$
Output: Metadata set $M_i$
1.    Preprocess Li (normalize, remove stopwords)
2.    Apply Named Entity Recognition (NER) for Urdu text
3.    Extract domain-specific keywords using frequency and context analysis
4.    Map extracted entities and keywords to DMS metadata fields
5.    Store $M_i$ in the document repository
```

Figure 5: Algorithmic workflow for metadata extraction

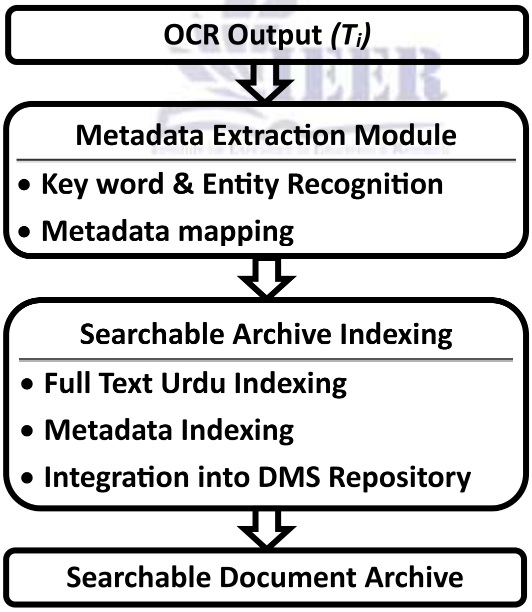
OCR Output (T_i)
Metadata Extraction Module
• Key word & Entity Recognition
• Metadata mapping
Searchable Archive Indexing
• Full Text Urdu Indexing
• Metadata Indexing
• Integration into DMS Repository
Searchable Document Archive

Figure 6: Accessible and searchable document archive

3.11 Experimental Setup

Experimental setup was implemented on the Google Colab platform and includes specification of hardware/software environment, training configuration, hyperparameters, loss functions, optimization strategy and evaluation metrics. Although the visible file excerpts do not expose all numerical settings, The experiments were conducted as controlled model training and evaluation under a defined computational environment.

3.12 Evaluation Metrics

The methodology used a multi-level evaluation framework consisting of OCR-level, language-model-level and system-level measures. The Character Error Rate (CER) and Word Error Rate (WER) were defined as the core OCR evaluation metrics. CER evaluates recognition quality at character level, while WER is more suitable for administrative documents where semantic correctness depends on word-level fidelity [13].

Equation (1): Character Error Rate (CER)

$$CER = \frac{S_c + D_c + I_c}{N_c}$$

where S_c, D_c and $I_c$denote the number of character-level substitutions, deletions and insertions, respectively and $N_c$is the total number of characters in the reference text [16], [17], [19].

Equation (2): Word Error Rate (WER)

$$WER = \frac{S_w + D_w + I_w}{N_w}$$

where S_w, D_w and $I_w$denote word-level substitutions, deletions and insertions, respectively and $N_w$is the total number of words in the reference text [16], [17], [19]..

Equation (3): Perplexity (PPL)

The perplexity was used to evaluate the predictive quality of the language model used in post-processing:

$$PPL(Y) = P(y_1, y_2, \dots, y_T)^{-\frac{1}{T}}$$

for a test sequence $Y = y_1, y_2, \dots, y_T$. Lower perplexity indicates better contextual prediction and stronger language-model support in post-processing [11], [12].

3.13 Ethical Considerations and Methodological Limitations

Because the study uses official and administrative documents, ethical considerations were explicitly included in the methodology. The privacy, confidentiality, anonymization of sensitive information, responsible AI use and controlled data access as key concerns [24], [25], [26]. The methodology also recognizes limitations related to low-resource language settings, scarcity of annotated data, domain-specific vocabulary and model precision under real-world document variability [7], [9].

3.14 Summary of the Methodology

In summary, the proposed methodology combined dataset preparation, preprocessing, segmentation strategy, deep learning-based OCR recognition, post-processing and DMS integration into a unified framework for Urdu administrative documents. The design was tailored specifically to the visual, linguistic and institutional characteristics of District Education Authority correspondence and provides a strong basis for evaluating OCR performance in a real educational governance setting.

4. Results and Discussion

This section analyses the results which are obtained after the experiments from the proposed smart OCR framework which was developed for administrative documents of DEAs which were written in Urdu. The analysis follows directly from the methodological pipeline described in the previous section and evaluates the system at multiple levels, including dataset characteristics, preprocessing performance, segmentation quality, recognition accuracy, post-processing gains, system-level efficiency and the comparative model behaviour. The purpose of this section is not only to report numerical outcomes, but also to interpret how the proposed framework performs under realistic document conditions involving noise, skew, variable layouts and cursive Nastaliq writing.

4.1 A. Dataset Analysis and Exploratory Findings

The dataset which was used for the purpose of evaluation consisted of multiple types of Urdu administrative documents, such as official letters, circulars, notifications and the reports. The frequency distribution shows that official letters formed the largest portion of the corpus, followed by circulars, notifications and reports. This distribution is important because it reflects the intended operational context of the proposed system, namely the digitization of real educational and administrative correspondence. A heterogeneous dataset of this kind improves the practical validity of the evaluation by exposing the OCR pipeline to diverse layouts, varying text densities and differences in print and scan quality [14], [21].

Table 4
Distribution of Urdu Document Types

Document Type	Number of Documents
Official Letters	450
Circulars	320
Notifications	280
Reports	150

The dataset analysis further indicates variability in font styles, writing conditions and image quality, including blurred scans, smudges, uneven lighting and skewed text lines. From a supervisory perspective, this is a strength rather than a limitation, because an OCR model intended for public-sector document processing must be evaluated on realistically imperfect material rather than artificially clean benchmarks [7], [8], [12]. The exploratory analysis therefore confirms that the dataset is suitable for examining the robustness and generalizability of the proposed OCR architecture.

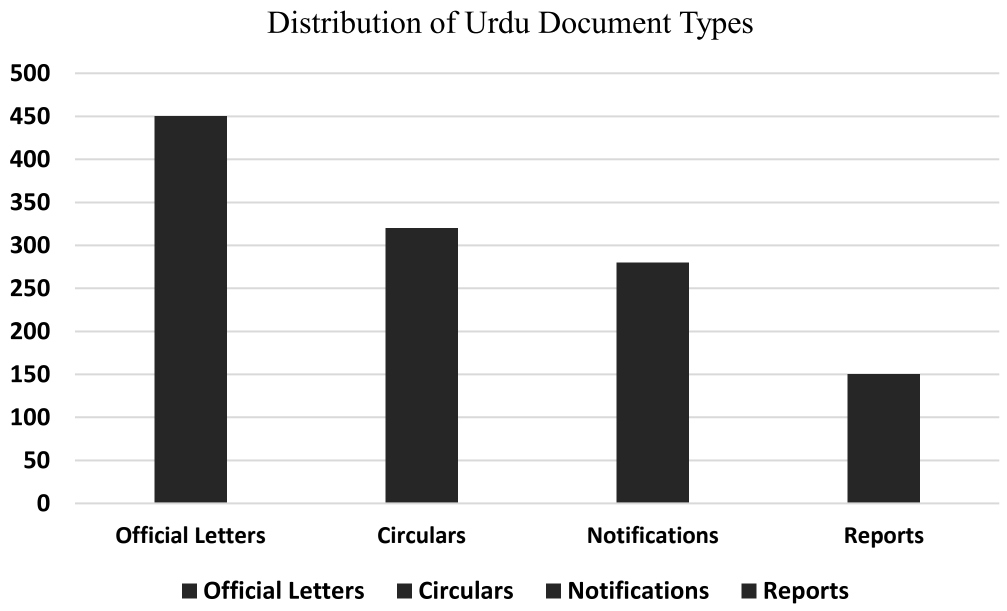

Figure 7: Distribution of Urdu documents types

4.2 Preprocessing Results

The preprocessing stage played a decisive role in improving the readability of scanned Urdu documents before recognition [7], [8]. Since Urdu Nastaliq script contains connected strokes, variable baselines and fine diacritics, image enhancement was necessary to reduce distortions without damaging character structure [2], [3], [8], [27]. The preprocessing pipeline included grayscale conversion, noise removal, binarization, skew detection and correction and layout normalization. Quantitative analysis shows that each step contributed to improving image quality and stabilizing the subsequent OCR stages. [28]

Adaptive thresholding proved more effective than global thresholding, particularly for older and degraded documents [22]. The adaptive binarization preserved strokes, ligatures and diacritics more effectively and reduced misclassification caused by partial character loss by 12%. This result is particularly relevant for Urdu OCR, because the loss of a small visual component may alter the entire word identity. In addition, Hough transform-based skew correction reduced initial skew angles in the range of ±10° to ±15° to a mean residual skew below 1°, leading to a 14% improvement in line segmentation accuracy and a 10% improvement in word segmentation consistency. These outcomes confirm that preprocessing was not merely cosmetic; it materially improved the quality of model input.

Table 5

Preprocessing Performance

Processing Step	**Noise Level Before (%)**	**Noise Level After (%)**
Grayscale Conversion	42	35
Noise Removal	35	18
Binarization	18	10
Skew Correction	10	5

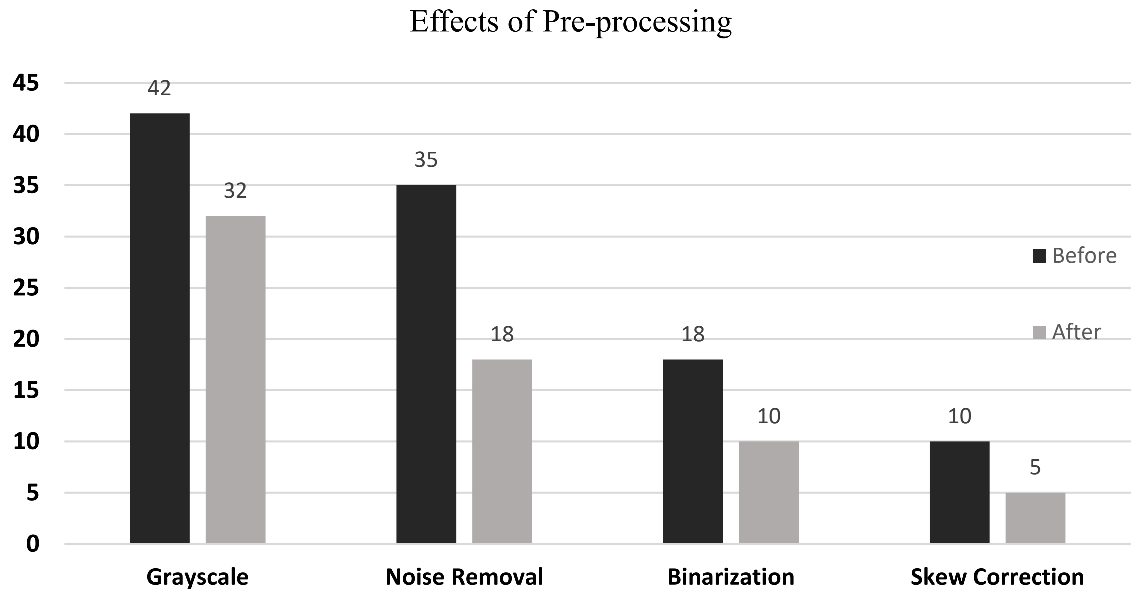

Figure 8: Effect of Preprocessing on Document Quality.

The results in Table 5 show a consistent reduction in document noise after each preprocessing stage, with the strongest improvement appearing during dedicated noise removal and binarization. These findings support the methodological decision to treat preprocessing as a central component of the OCR pipeline rather than as a minor preparatory step.

For completeness, the evaluation metrics used to interpret downstream recognition performance remain the standard OCR measures of character error rate and word error rate:

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

where $\boldsymbol{S}$ denotes substitutions, $\boldsymbol{D}$ deletions, $\boldsymbol{I}$ insertions and $\boldsymbol{N}$ the total number of reference characters.

$$\boldsymbol{WER} = \frac{\boldsymbol{S_w} + \boldsymbol{D_w} + \boldsymbol{I_w}}{\boldsymbol{N_w}}$$

where $\boldsymbol{S_w}$, $\boldsymbol{D_w}$ and $\boldsymbol{I_w}$refer to word-level substitutions, deletions and insertions, while $\boldsymbol{N_w}$represents the total number of words in the reference text.

4.3 Segmentation Performance

Segmentation performance was evaluated at the line, word and ligature levels. This stage is especially challenging in Urdu Nastaliq because the script is cursive, diagonally flowing and structurally dependent on context-sensitive joining patterns [28], [29]. The line segmentation achieved high reliability, with an accuracy of 95% on moderately sloped text and 91% on highly curved lines. These results indicate that the adopted layout normalization strategy successfully reduced the impact of skew and baseline variation.

However, segmentation at the word and ligature levels remained more difficult [19], [22], [30]. Overlapping strokes, variable spacing and the attachment of diacritics often complicate the separation of connected text units [10], [11], [19]. This finding is consistent with the broader OCR literature for cursive scripts and also supports the architectural decision to rely on sequence-based and Transformer-assisted recognition rather than strict character-level segmentation. In supervisory terms, this is a sound design choice, instead of forcing brittle segmentation rules onto a complex script, the system shifts part of the burden to models capable of learning contextual dependencies directly from visual sequences [28], [31], [33].

4.4 OCR Recognition Results

The recognition stage examined the contribution of CNN-based feature extraction, sequence modeling and Transformer-based decoding. Training performance improved steadily across epochs, indicating effective convergence and stable learning behavior [8], [17]. The training accuracy increased from 55% in the first epoch to 96% by epoch 30, while validation accuracy increased from 52% to 94% over the same training span. The relatively narrow gap between training and validation performance suggests that the model generalized reasonably well without severe overfitting.

Table 6
OCR Training Accuracy vs. Epochs

Epoch	Training Accuracy (%)	Validation Accuracy (%)
1	55	52
5	68	64
10	78	74
15	86	83

20	91	88
25	94	92
30	96	94

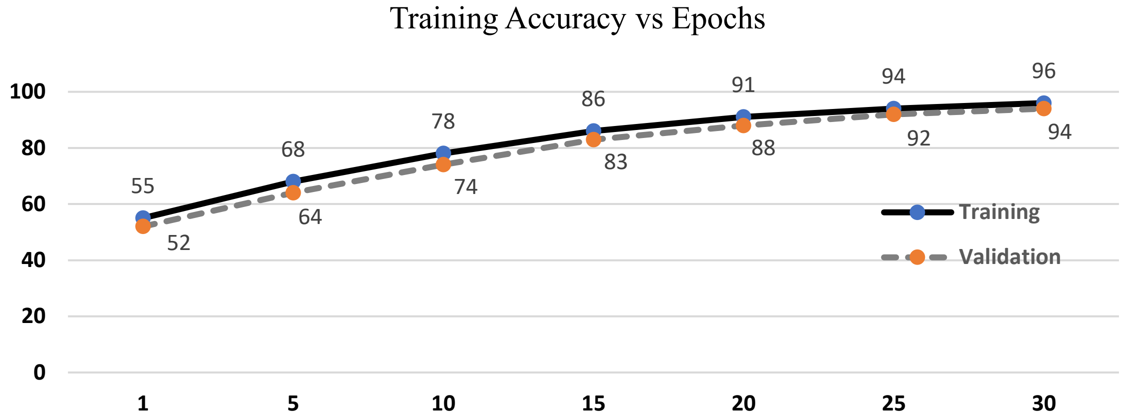

Figure 9: OCR Training Accuracy vs Epochs

The proposed frame work shows the strong performance and good results on the selected dataset. The reported system-level metrics include 95.8% character accuracy and 92.6% word accuracy, with a CER of 4.2% and a WER of 7.4%. The results show that the proposed model preserved the detailed character recognition and it maintained the strong word-level coherence for archiving and retrieval of data. For a low-resource, cursive and context-sensitive language such as Urdu, these values represent a strong result and support the viability of the proposed architecture for administrative digitization [7], [9].

Table 7
OCR Performance Metrics

Metric	Value (%)
Character Accuracy	95.8
Word Accuracy	92.6
Character Error Rate (CER)	4.2
Word Error Rate (WER)	7.4

4.5 Post-Processing and Language Modeling Results

The post-processing contributed meaningful improvements beyond the raw recognition output. Dictionary-based correction reduced character-level errors by 5% to 7% and word-level errors by 6% to 8%, indicating that lexicon-guided refinement was effective in recovering visually ambiguous terms. Rule-based post-processing produced an additional 3% to 4% reduction in recurring errors, especially those related to ligature formation and diacritic placement. These results are important because they show that recognition quality in Urdu OCR should not be assessed at the visual level alone; language-aware refinement remains essential for producing usable final text.

Neural language model integration generated further gains in contextual correctness. The LSTM-based language model reported a perplexity of 21.5, whereas the Transformer-based language model achieved a lower perplexity of 18.2, indicating stronger contextual modeling. It was also found and later confirmed that during this working an overall contextual accuracy improvement of 4% to 6% was achieved after integration of the neural language model. From a supervisory standpoint, this part of the study is especially valuable because it moves the contribution beyond image recognition into linguistically meaningful text recovery, which is what real institutional workflows actually require.

4.6 System-Level Evaluation

At the system level, the OCR framework was evaluated in terms of accuracy, throughput, scalability and robustness [9], [13], [29]. The pipeline processed approximately three to four pages per second on a GPU-based environment, suggesting that the system is practical for medium- and large-scale digitization tasks. It is further noted that the model maintained stable performance across datasets with variations in size, quality, skew and font style. This matters because public-sector archives typically contain mixed collections rather than uniform document sets [13].

The system-level analysis also reinforces the reliability of the recognition pipeline under realistic degradation conditions. Moderate level of noise and skew reduced the WER by less than 2% and show that preprocessing ‹sequence recognition and the contextual correction created a robust in workflow of the OCR. The overall system can therefore be understood not merely as an OCR model, but as a deployable document-processing pipeline.

References

[1] Yasin, M., & Gondhi, N.K. (2020). A Comparative Analysis on Nastaliq Style Urdu Character Recognition. Journal of Computational and Theoretical Nanoscience, 17, 284-289.

[2] Kazmi, M., Yasir, F., Habib, S.M., Hayat, M., & Qazi, S.A. (2021). Photometric Ligature Extraction Technique for Urdu Optical Character Recognition. Engineering, Technology & Applied Science Research.

[3] Shabbir, S., Javed, N., Siddiqi, I., & Khurshid, K. (2017). A comparative study on clustering techniques for Urdu ligatures in nastaliq font. 2017 13th International Conference on Emerging Technologies (ICET), 1-6.

[4] Naseer, A., & Zafar, K. (2022). Meta-feature based few-shot Siamese learning for Urdu optical character recognition. Computational Intelligence, 38, 1707 - 1727.

[5] R. Smith, "An overview of the Tesseract OCR engine," Proc. International Conference on Document Analysis and Recognition (ICDAR), pp. 629–633, 2007.

[6] A. Graves, M. Liwicki, S. Fernandez, R. Bertolami, H. Bunke and J. Schmidhuber, "A novel connectionist system for unconstrained handwriting recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 31, no. 5, pp. 855–868, 2009.

[7] A. Mustafa, S. Malik and H. Javed, "Transformer-based OCR for cursive and low-resource scripts," International Journal of Computer and Informatics, vol. 10, no. 3, pp. 45–58, 2024.

[8] Naeem, M.F., Zia, N.U., Awan, A.A., Shafait, F., & Ul-Hasan, A. (2017). Impact of Ligature Coverage on Training Practical Urdu OCR Systems. 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), 01, 131-136.

[9] A. Ul-Hasan et al., "Evaluation metrics for OCR accuracy in low-resource languages," International Journal of Computer Science and Engineering, 2022.

[10] S. Faizullah, M. S. Ayub, S. Hussain and M. A. Khan, "A survey of OCR in Arabic language: Applications, techniques and challenges," Applied Sciences, vol. 13, no. 7, 2023.

[11] M. G. Mahdi, A. Sleem, I. M. Elhenawy and S. Safwat, "Towards robust Arabic and Urdu OCR systems: A systematic review of deep learning techniques," International Journal of Computer and Informatics, 2023.

[12] A. Rahman, A. Ghosh and C. Arora, "UTRNet: High-resolution Urdu text recognition in printed documents," Proc. ICDAR, 2023.

[13] A. Mustafa, M. Shahzad and K. Saghar, “UTRSet: Annotated Urdu text datasets for OCR evaluation,” arXiv preprint arXiv:2408.15119, 2024.

[14] T. Anjum and A. Azhar, “A survey on Urdu handwritten text recognition: State of the art, challenges and future directions,” Journal of Computing and Artificial Intelligence, 2025.

[15] Zahid, H., Rashid, M., Hussain, S., Azim, C.F., Syed, S.A., & Saad, A. (2022). Recognition of Urdu sign language: a systematic review of the machine learning classification. PeerJ Computer Science, 8.

[16] Akram, Q.U., & Hussain, S. (2019). Improving Urdu Recognition Using Character-Based Artistic Features of Nastalique Calligraphy. IEEE Access, 7, 8495-8507.

[17] Ganai, A.F., & Khursheed, F. (2023). Computationally efficient recognition of unconstrained handwritten Urdu script using BERT with vision transformers. Neural Computing and Applications, 35, 24161-24177.

[18] Yasir, F., & Kazmi, M. (2025). Acceleration of Urdu Optical Character Recognition on Zynq UltraScale+ MPSoC Using Deep Convolutional Neural Network. IEEE Access, 13, 135538-135557.

[19] Khan, H.R., Kazmi, M., Khalid, H., Hasan, M.A., Fayyaz, N., Ahmed, S., & Qazi (2021). A Holistic Approach to Urdu Language Word Recognition using Deep Neural Networks. Engineering, Technology & Applied Science Research.

[20] Narwani, K., Lin, H., Pirbhulal, S., & Hassan, M. (2025). Toward AI-Enabled Approach for Urdu Text Recognition: A Legacy for Urdu Image Apprehension. IEEE Access, 13, 122022-122034.

[21] Saeed, A., & Qamar, F. (2020). Improving Nastaliq OCR Accuracy Using Data Augmentation. International Journal of Pattern Recognition and Artificial Intelligence.

[22] M.A., Khan, H., Khan, F.A., Kumar, K., Wagan, A.A., & Solangi, S. (2020). Handwritten Urdu character recognition via images using different machine learning and deep learning techniques. Indian journal of science and technology, 13, 1746-1754

[23] Ahmed, T., & Butt, M.U. (2023). Recognition of Urdu Handwritten Words Using Deep Learning Techniques. 2023 20th International Bhurban Conference on Applied Sciences and Technology (IBCAST), 261-266.

[24] Tayyab, M., Hussain, A., Alshara, M.A., Khan, S., Alotaibi, R.M., & Baig, A.R. (2022). Recognition of Visual Arabic Scripting News Ticker from Broadcast Stream. IEEE Access, 10, 59189-59204.

[25] Rashid, A., Aslam, A., & Baig, M. (2021). Challenges in Urdu Nastaliq OCR: A Comprehensive Review. Journal of Intelligent Systems.

[26] Rashid, A., Mahmood, S., Inayat, U., & Zia, M.F. (2025). Urdu Toxicity Detection: A Multi-Stage and Multi-Label Classification Approach. AI.

[27] Rafeeq, M., ur Rehman, Z., Khan, A., Khan, I.A., & Jadoon, W. (2018). Ligature categorization based Nastaliq Urdu recognition using deep neural networks. Computational and Mathematical Organization Theory, 25, 184 - 195.

[28] Naseer, A., Hussain, S., Zafar, K., & Khan, A. (2021). A novel normal to tangent line (NTL) algorithm for scale invariant feature extraction for Urdu OCR. International Journal on Document Analysis and Recognition (IJDAR), 25, 51 - 66.

[29] Hassan, S., Raza Shahid, A., & Asif Naeem, M. (2025). TDA-ViT: A Transformer-Based Framework for Unified Urdu Text Recognition via Topological and Visual Feature Fusion. IEEE Access, 13, 182940-182959.

[30] Tayyab, M., & Hussain, A. (2023). Convolutional Matching Technique for Urdu Text Recognition. 2023 25th International Multitopic Conference (INMIC), 1-5.

[31] Rehman, A.U., & Hussain, S.U. (2020). Large Scale Font Independent Urdu Text Recognition System. ArXiv, abs/2005.06752.

[32] Ahmed, G.S., Alyas, T., Waseem Iqbal, M., Usman Ashraf, M., Mohammed Alghamdi, A., A. Bahaddad, A., & Ali Almarhabi, K. (2022). Recognition of Urdu Handwritten Alphabet Using Convolutional Neural Network (CNN). Computers, Materials & Continua.

[33] S. Saber and M. G. Mahdi, "Urdu handwriting recognition with deep learning: Current methods and future prospects," International Journal of Computer and Informatics, 2024.