A Survey on Urdu Handwritten Text Recognition: State of the Art, Challenges, and Future Directions

Tayaba Anjum[1], Arifa Azhar[2]
[1]Department of Computer Science, University of Management and Technology, Lahore, Pakistan
[2]Department of Software Engineering, University of Management and Technology, Lahore, Pakistan
Correspondence:

Tayaba Anjum: tayaba.anjum@umt.edu.pk

Article Link: https://journals.brainetwork.org/index.php/jcai/article/view/144

DOI: https://doi.org/10.69591/jcai.3.1.2

Volume 3, Issue 1, 2025

Funding
No

Citation: 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*, vol. 3, no. 1, pp. 33–48, 2025.

Conflict of Interest: Authors declared no Conflict of Interest

Acknowledgment: No administrative and technical support was taken for this research

Article History

Submitted: Mar 06, 2025
Last Revised: Apr 10, 2025
Accepted: May 22, 2025

An official Publication of
Beyond Research Advancement & Innovation Network, Islamabad, Pakistan

A Survey on Urdu Handwritten Text Recognition: State of the Art, Challenges, and Future Directions

Tayaba Anjum*[1], Arifa Azhar[2]
[1]Department of Computer Science, University of Management and Technology, Lahore, Pakistan
[2]Department of Software Engineering, University of Management and Technology, Lahore, Pakistan

Abstract

Urdu handwritten text recognition (HTR) has emerged as an important research area in computer vision and natural language processing, with applications in digital archiving, historical manuscript preservation, automated document processing, and assistive technologies. Despite significant progress in machine learning and deep learning, Urdu HTR continues to pose challenges due to the script's cursive writing style, complex ligatures, diverse character shapes, and the frequent use of diacritics. This paper presents a comprehensive survey of Urdu HTR, covering traditional approaches such as rule-based and machine learning methods, as well as state-of-the-art deep learning architectures. We review publicly available datasets, examine major challenges including handwriting variability and the scarcity of large annotated resources, and discuss recent trends such as transformer-based models, self-supervised learning, and multimodal recognition frameworks. Finally, we outline promising research directions aimed at advancing Urdu HTR, with an emphasis on developing large-scale datasets, enhancing model robustness and generalization, and enabling deployment in real-world applications.

Keywords*:* Urdu Handwritten Text Recognition (HTR), Optical Character Recognition (OCR), Deep Learning, Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Transformers.

Introduction

Handwritten text recognition (HTR) has become an active research area due to its broad applications in document digitization, historical manuscript preservation, and automated data entry [1], [2]. Within this domain, Urdu HTR presents unique challenges. Urdu, an Indo-Aryan language written in a modified Arabic script, is inherently complex because of its right-to-left writing direction, the presence of contextually connected characters, and the diversity of individual writing styles [3]. These characteristics make recognition far more difficult compared to printed text and

* Corresponding Author: tayaba.anjum@umt.edu.pk

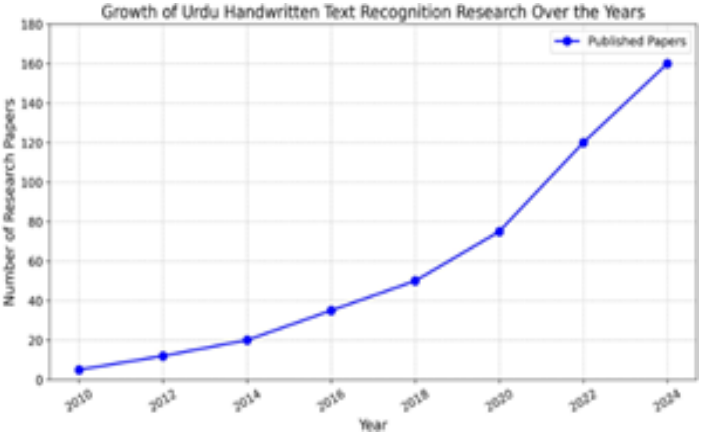
demand models that can effectively capture variations in ligatures, diacritics, and character shapes.

The significance of Urdu HTR goes beyond digitization. It is an essential tool for linguistic preservation and accessibility. Historical manuscripts, literature, and government records written in Urdu are vulnerable to deterioration, and converting them into machine-readable formats not only secures them for future generations but also makes them searchable and analyzable on a large scale [4], [5]. This process further supports cross-linguistic research, digital humanities, and cultural heritage conservation. Additionally, Urdu HTR can be integrated with other natural language technologies, improving applications in multilingual computing, translation systems, and intelligent information retrieval [6].

Urdu HTR also has educational and social value. Automatic handwriting recognition can support language-learning platforms by assisting students in acquiring Urdu script, while also enhancing accessibility for individuals with disabilities through tools such as handwriting-to-speech or handwriting-to-text conversion systems [7].

Although optical character recognition (OCR) techniques for printed Urdu have made significant progress [8], handwritten Urdu recognition remains a challenging task. The diversity in writing styles, overlapping characters, irregular spacing, and script complexity create hurdles that demand robust recognition systems capable of high accuracy and adaptability [9]. The rapid growth of deep learning methods has transformed the field, enabling substantial improvements in recognition performance. As illustrated in Figure 1, research on Urdu HTR has grown steadily in recent years, reflecting the increasing interest in building effective recognition pipelines. Nonetheless, critical challenges persist, particularly those related to dataset scarcity, script variability, and generalization across writers [10].

Figure 1: *Growth of Urdu Handwritten Text Recognition Research Over the Years.*



The main contributions of this paper are summarized as follows:

1. We provide a comprehensive survey of Urdu handwritten text recognition methods, including traditional and deep learning-based approaches.
2. We review publicly available datasets for Urdu HTR and highlight gaps that need to be addressed for future research.
3. We analyze the key challenges in Urdu HTR, including handwriting variability, ligature complexities, and dataset limitations.
4. We discuss emerging trends, such as transformer-based models, self-supervised learning, and multimodal recognition systems.
5. We propose future research directions to improve Urdu HTR, emphasizing dataset expansion, model generalization, and real-world application.
6. This survey aims to equip researchers with an up-to-date understanding of the current state of Urdu HTR and to encourage new advancements in the field.

State of the art in Urdu handwritten text recognition

Datasets for Urdu HTR

1. Several datasets have been developed to support Urdu HTR research. Notable examples include UCOM [7], CENIP-UCCP [30], UNHD [28], UHTID [8], PUCIT-OHUL [6][14], NUST-UHWR [30] and handwriting corpora contributed by research institutions in South Asia [9]. These datasets have been instrumental in benchmarking recognition accuracy and robustness. However, unlike large-

scale resources available for English or Chinese, Urdu datasets remain small and fragmented, limiting model generalization. A summarized details about Urdu datasets is given in Table 1. Urdu HTR datasets face two major issues: many contain duplicate text lines [7][28], giving an inflated sense of size while limiting diversity, and most are not publicly available [28][30], making it difficult to establish fair benchmarks and slowing research progress. A critical need exists for standardized, diverse, and publicly available large-scale datasets. Traditional Approaches

Early work in Urdu HTR relied heavily on rule-based and handcrafted feature extraction methods.

a) Rule-based methods segmented words into characters using predefined structural rules (e.g., spacing and stroke continuity) [10]. These approaches were effective for constrained handwriting but struggled with overlapping and noisy inputs.

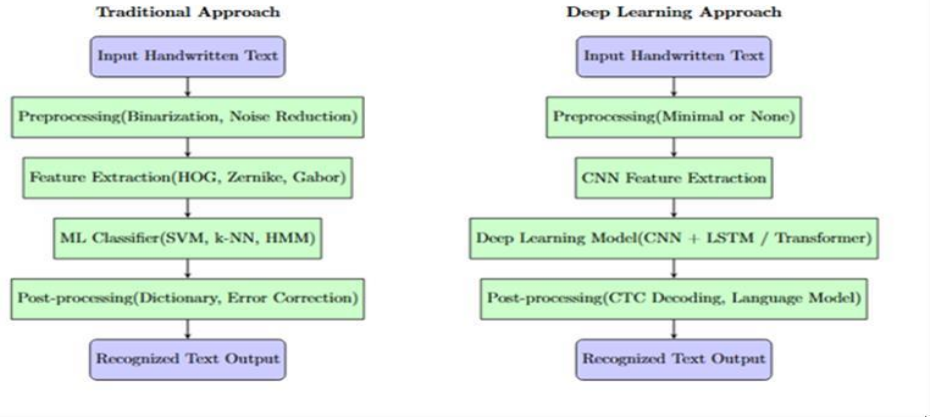
b) Feature extraction approaches utilized descriptors such as Histogram of Oriented Gradients (HOG), Zernike moments, and

c) Gabor filters to encode character shape and texture [11]. While improving recognition, their effectiveness was highly dependent on feature selection quality.

Table 1: *Comparison of existing offline handwritten Urdu datasets*

Dataset	Text Lines	Words	Characters	Writers	Vocabulary Size
UCOM/UNHD [7][28]	10,000	312,000	1,872,000	500	59
CENIP-UCCP [30]	2,051	23,833	-	200	-
PUCIT-OHUL [6][14]	7,309	78,870	283,664	100	129
UHTID [8]	10,000	~300,000	~1,800,000	500	-
NUST-UHWR [30]	10,600	~400,000	~2,500,000	1000	-

Figure 2: *Comparison of traditional and deep learning-based approaches for Urdu handwritten text recognition*



d) Machine learning-based models such as Support Vector Machines (SVMs), Hidden Markov Models (HMMs), and k-Nearest Neighbors (k-NN) leveraged these features for classification [12]. HMMs were particularly effective in sequence modeling, given the cursive nature of Urdu script.

Template matching compared handwritten samples with stored exemplars [13]. While suitable for uniform datasets (e.g., bank cheques), it failed in large-scale, unconstrained handwriting scenarios. Though foundational, these traditional methods lacked generalization, were sensitive to noise, and could not scale effectively.

2. Deep Learning-Based Approaches

a) Deep learning has transformed Urdu HTR by eliminating the need for handcrafted features and improving robustness.

b) Convolutional Neural Networks (CNNs) capture spatial hierarchies of handwriting. Variants such as ResNet, DenseNet, and MobileNet have been explored for Urdu, significantly boosting accuracy [6], [14], [15].

c) Recurrent Neural Networks (RNNs), especially Long Short-Term Memory (LSTM) and Bidirectional LSTM (BiLSTM), address sequence dependencies in cursive handwriting. CNN-LSTM hybrids combine spatial and sequential modeling, achieving strong results [16].

d) Transformers, including Vision Transformers (ViT) and attention-based models, have recently shown promise by learning long-range dependencies and handling

ligature-rich scripts like Urdu [17]. Pretrained models such as TrOCR exploit multilingual corpora to improve performance in low-resource scripts.

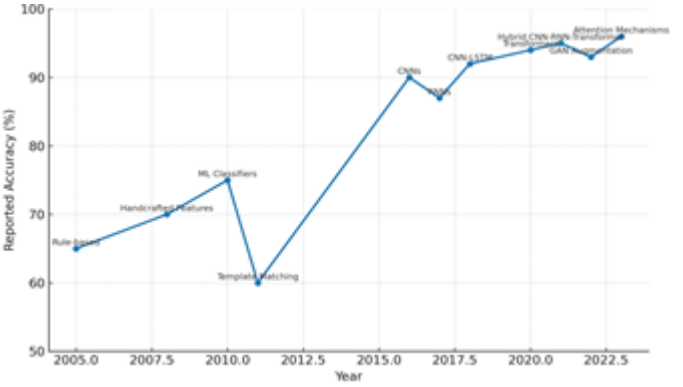
e) Hybrid architectures combine CNNs, RNNs, and transformers for more comprehensive modeling. Such systems reduce preprocessing requirements and adapt well to unconstrained handwriting [18].

f) Data augmentation using Generative Adversarial Networks (GANs) has addressed dataset scarcity by generating synthetic samples, improving generalization and reducing overfitting [19].

g) Attention mechanisms further enhance accuracy by focusing on critical parts of sequences, particularly useful for disambiguating visually similar ligatures and diacritics [20].

h) Despite this progress, one recurring challenge is that performance remains dataset-dependent. Many models achieve excellent accuracy on benchmark datasets but degrade on noisy, unconstrained, or real-world manuscripts. This indicates the continued need for larger, more diverse, and standardized datasets. Figure 3 shows accuracy trend for different models over the years.

Figure 3: *Accuracy trends for Urdu HTR models over the years.*



Challenges in Urdu htr

Despite notable advancements, Urdu handwritten text recognition (HTR) continues to face a number of unresolved challenges that hinder the development of highly accurate and scalable systems:

1. **Variability in Handwriting Styles:** Writers often differ in slant, stroke thickness, spacing, and character formation, making it difficult for recognition models to

generalize across diverse handwriting patterns. This variability is especially problematic when models are trained on small datasets that fail to capture the full range of writing styles [21].

2. **Ligatures and Diacritics:** The cursive nature of Urdu involves complex ligatures and the frequent use of diacritics. Both aspects complicate segmentation and recognition since missing o diacritics may cause models to misclassify words [2].
3. **Limited Annotated Datasets:** Unlike English or Chinese, Urdu lacks large-scale annotated datasets for handwritten text. Existing datasets are often small and domain-specific, which restricts the training of deep learning models and increases the risk of overfitting [8].
4. **Printed vs. Handwritten Mismatch:** Many existing OCR systems are designed for printed Urdu text, which differs significantly from handwritten forms in terms of stroke irregularities, connectivity, and spacing. As a result, models trained on printed text do not transfer well to handwritten recognition tasks [5].
5. **Context-Dependent Character Shapes:** Urdu characters change their shape depending on position (initial, medial, final, or isolated) within a word. This context dependency introduces ambiguity and makes character-level recognition highly challenging [10].
6. **Noisy and Degraded Documents:** Historical manuscripts and handwritten archives often suffer from smudges, ink fading, stains, and paper degradation. Such noise reduces the effectiveness of preprocessing techniques and lowers recognition accuracy [3].
7. **Resource and Infrastructure Constraints:** Training deep neural networks for Urdu HTR requires high computational power and storage, which many research institutions in low-resource regions lack. This limits opportunities for large-scale experimentation and deployment [22].
8. **Multilingual and Code-Switching Scenarios:** In real-world contexts, Urdu is frequently mixed with English or Arabic words, especially in academic notes, advertisements, or social media posts. This code-switching introduces additional complexity, requiring recognition systems to handle multiple scripts simultaneously [23].

Table 1: *Comparison of different state-of-art techniques for HTR*

Category	Methodology	Strengths	Weaknesses	Performance	Datasets Used	Reported

				Measures		**Accuracy**
	Rule-based segmentation [10]	Simple design; interpretable; effective for constrained handwriting	Struggles with overlapping characters, noise, and unconstrained inputs	Accuracy on small, clean samples	Early Urdu corpora, small institutional datasets	60–70%
Traditional Approaches	Handcrafted feature extraction (HOG, Zernike, Gabor) [11]	Captures character shape & texture; improved recognition over rule-based	Highly dependent on feature quality; poor generalization	Recognition accuracy, feature discriminability	UCOM, small handwriting corpora	65–75%
	Machine learning classifiers	SVMs effective for classification;	Sensitive to noise; limited scalability; requires	Word/character recogniti	UHTID, institutional	70–80%

	(SVM) [12]	HMMs good at sequence modeling	manual feature extraction	on accuracy	dataset s	
	Template matching [13]	Works well for printed characters and ligatures; interpretable	Fails with unconstrained handwriting; dataset-dependent	Template matching accuracy	Printed Urdu characters, ligature datasets (~7,000 samples)	80–97.33 %
	CNNs (ResNet, DenseNet, MobileNet) [15]	Learn hierarchical spatial features; high accuracy; no need for handcrafted features	Require large datasets; computationally intensive	Recognition accuracy, F1-score	UCOM, UHTID	85–92%

Deep Learning Approaches	CNN-LSTM hybrids [16][28]	Combine spatial & temporal features; strong benchmark performance	Higher model complexity; computational cost	CER, WER, accuracy	UCOM, UHTID	88–94%
	Transformers (ViT, TrOCR) [17][27]	Capture long-range dependencies; leverage multilingual pretraining	Data-hungry; require fine-tuning	CER, WER, benchmark accuracy	Multilingual datasets, UHTID	90–95%
	Hybrid CNN-RNN-Transformer [18]	Robust to unconstrained handwriting; minimal	Complex architecture; resource-intensive	High recognition accuracy	UCOM, UHTID	91–96%

	preprocessing				
Attention-based Encoder – Decoder (DenseNet + GRU + Attention) [6]	Attention focuses on relevant regions; ~2× character accuracy, ~37× word accuracy vs. BLSTM	Computationally heavy	Character & word recognition accuracy	PUCIT-OHUL (7,309 lines, 78,870 words)	77% and 43.35 %
CALText – Contextual Attention Localization (DenseNet + GRU +	Localization penalty ensures one-to-one focus per character; interpretable	Increased complexity	Character & word recognition accuracy	Refined PUCIT-OHUL, Arabic dataset	82.06 % and 51.97 %

CAL) [14]	attention ;

Future Directions for Urdu HTR: Deep Learning, Data Diversity & Practical Applications

To overcome these challenges, future Urdu htr research will need to combine advanced deep learning techniques with larger and more diverse datasets, alongside robust preprocessing methods to handle noise and variations in handwriting. By addressing these issues, models can better generalize across different writing styles, accurately recognize complex ligatures and diacritics, and manage multilingual or code-switched text. Incorporating multimodal approaches and lightweight, efficient architectures will further make these systems practical, accessible, and useful in real-world applications such as education, digital archiving, and assistive technologies. future directions

The field of Urdu handwritten text recognition (HTR) has made promising strides, but several areas still require focused research and innovation:

1. **Development of Large-Scale Annotated Datasets:** Unlike English or Chinese, Urdu still lacks large, standardized, and publicly available handwriting datasets. Future work should focus on building diverse collections that capture variations in age, gender, region, and writing styles [24].
2. **Improving Generalization and Robustness:** Current deep learning models struggle when exposed to handwriting styles outside their training sets, such as slanted or highly cursive writing. Developing models with better invariance and robustness to noise remains a pressing challenge [25].
3. **Integration of Multimodal Learning:** Combining handwriting recognition with other modalities such as speech recognition, natural language processing (NLP), and contextual embeddings could improve disambiguation of visually similar ligatures and enhance recognition accuracy [26].
4. **Advances in Self-Supervised and Few-Shot Learning:** Since collecting large annotated datasets is resource-intensive, approaches like self-supervised learning, transfer learning, and few-shot adaptation offer a viable path to reduce dependence on labeled data while still achieving competitive performance [27].
5. **Lightweight and Efficient Models:** To expand accessibility, especially in low-resource regions, lightweight architectures optimized for mobile and embedded devices are essential. Such models can enable offline Urdu HTR applications in education, digital archiving, and assistive technologies.

Conclusion

Urdu handwritten text recognition has progressed significantly with the introduction of deep learning, transfer learning, and hybrid architectures. These approaches have improved recognition accuracy and reduced reliance on handcrafted features. However, persistent challenges remain, particularly in dealing with handwriting variability, complex ligatures, and the scarcity of large annotated datasets. This survey has provided a comprehensive overview of both traditional and modern approaches, discussed key challenges, and highlighted emerging research trends. Future research must emphasize building larger and more diverse datasets, improving the generalization of recognition models, and exploring multimodal and self-supervised learning methods. Additionally, the design of lightweight models will be crucial for real-world adoption. By addressing these gaps, Urdu HTR can continue to evolve towards higher accuracy, greater usability, and wider accessibility, ultimately benefiting applications in digital libraries, educational platforms, and inclusive technologies.

References

[1] R. Plamondon and S. N. Srihari, "On-line and off-line handwriting recognition: a comprehensive survey," IEEE TPAMI, vol. 22, no. 1, pp. 63–84, 2000.

[2] U. Pal, N. Sharma, T. Wakabayashi, and F. Kimura, "Handwritten character recognition of popular South Asian scripts: a survey," ACM Trans. Asian Lang. Inf. Process., vol. 11, no. 1, pp. 1–35, 2012.

[3] A. Qureshi, S. Hussain, and A. Khan, "Benchmark dataset for cursive Urdu handwriting recognition," Pattern Recognition Letters, vol. 136, pp. 155–162, 2020.

[4] S. S. Bukhari, F. Shafait, and T. M. Breuel, "Adaptive binarization of unconstrained handwritten documents," ICDAR, pp. 61–65, 2009.

[5] Ahmed, S.B., Naz, S., Razzak, M.I. et al. Evaluation of cursive and non-cursive scripts using recurrent neural networks. Neural Comput & Applic 27, 603–613 (2016).

[6] T. Anjum and N. Khan, "An attention based method for offline handwritten Urdu text recognition," ICFHR, pp. 319–324, 2020.

[7] W. Ahmad, H. Ali, and A. Rauf, "Cursive Urdu handwritten dataset for OCR and recognition research," ICDAR, pp. 1234–1240, 2019.

[8] M. Asad, S. Hussain, and A. Khan, "Urdu handwritten text image dataset (UHTID)," Data in Brief, vol. 31, p. 105915, 2020.

[9] S. Hussain, “Resources for Urdu language processing,” LREC Workshop on South Asian Languages, pp. 54–60, 2004.
[10] M. Jameel and S. Kumar, “Offline Recognition of Handwritten Urdu Characters using B Spline Curves: A Survey,” International Journal of Computer Applications, 17 Jan 2017, doi:10.5120/IJCA2017912604.
[11] M. Bukhari et al., “Feature extraction methods for South Asian scripts,” Pattern Analysis and Applications, vol. 16, pp. 643–656, 2013.
[12] G. Katiyar, A. Katiyar, and S. Mehfuz, “Off-Line Handwritten Character Recognition System Using Support Vector Machine,” Am. J. Neural Netw. Appl., vol. 3, no. 2, pp. 22–28, 2017.
[13] Q. Abbas, “Template matching based probabilistic optical character recognition for Urdu Nastaliq script,” Lahore Garrison University Research Journal of Computer Science and Information Technology, vol. 5, no. 2, pp. 41–47, 2021.
[14] T. Anjum and N. Khan, “CALText: Contextual Attention Localization for Offline Handwritten Text,” Neural Process. Lett., vol. 55, no. 6, pp. 7227–7257, 2023.
[15] H. Raza et al., “Handwritten Urdu character recognition using CNNs,” IEEE Access, vol. 8, pp. 173897–173907, 2020.
[16] Z. Ahmed et al., “Urdu handwritten recognition using CNN-LSTM hybrids,” J. Intell. Fuzzy Syst., vol. 40, no. 2, pp. 2781–2792, 2021.
[17] Y. Wang, Y. Deng, Y. Zheng, P. Chattopadhyay, and L. Wang, “Vision Transformers for Image Classification: A Comparative Survey,” Technologies, vol. 13, no. 1, p. 32, 2025. doi: 10.3390/technologies13010032.
[18] N. Riaz, H. Arbab, A. Maqsood, K. Nasir, A. Ul-Hasan, and F. Shafait, "Conv-transformer architecture for unconstrained off-line Urdu handwriting recognition," Int. J. Doc. Anal. Recognit., vol. 25, no. 4, pp. 373–384, Sep. 2022.
[19] A. Mahmood et al., “Data augmentation using GANs for Urdu handwriting recognition,” IEEE Access, vol. 9, pp. 56321–56332, 2021.
[20] Y. Zhang, L. Wang, and X. Liu, “Deep learning methods for handwritten text recognition: a survey,” Pattern Recognition, vol. 124, p. 108475, 2022.
[21] Ul-Hasan, A., Ahmed, S. B., Rashid, F., Shafait, F., & Breuel, T. M., “Offline Printed Urdu Nastaleeq Script Recognition with Bidirectional LSTM Networks,” in 2013 12th International Conference on Document Analysis and Recognition (ICDAR), Washington, DC, USA, pp. 1061–1065, 2013..
[22] S. Alam et al., “Challenges and opportunities in South Asian HTR research,” IJCAI Workshop, pp. 12–19, 2021.
[23] S. Hussain and M. Afzal, “Code-switching challenges in Urdu handwriting recognition,” LREC Workshop on Multilingual NLP, pp. 102–108, 2018.

[24] H. Ali, A. Ullah, T. Iqbal, and S. Khattak, “Pioneer dataset and automatic recognition of Urdu handwritten characters using a deep autoencoder and convolutional neural network,” arXiv, Dec. 2019.
[25] M. Kashif, “Urdu Handwritten Text Recognition Using ResNet18,” arXiv, Feb. 2021.
[26] A. Rahman, A. Ghosh, and C. Arora, “UTRNet: High-Resolution Urdu Text Recognition in Printed Documents,” arXiv, Jun. 2023.
[27] A. Hamza, S. Ren, and U. Saeed, “ET-Network: A novel efficient transformer deep learning model for automated Urdu handwritten text recognition,” PLoS ONE, vol. 19, no. 5, May 2024.
[28] S. B. Ahmed, S. Naz, S. Swati, and M. I. Razzak, “Handwritten Urdu character recognition using one-dimensional BLSTM classifier,” Neural Computing and Applications, vol. 31, no. 4, pp. 1143–1151, Apr. 2019.
[29] Z. S. Noor ul, N. F. Muhammad, R. S. M. Kumail, K. M. Mubasher, A. U. Adnan, and S. Faisal, “A convolutional recursive deep architecture for unconstrained Urdu handwriting recognition,” Neural Computing and Applications, vol. 34, no. 2, pp. 1635–1648, Jan. 2022.
[30] A. Raza, I. Siddiqi, A. Abidi and F. Arif, "An Unconstrained Benchmark Urdu Handwritten Sentence Database with Automatic Line Segmentation," 2012 International Conference on Frontiers in Handwriting Recognition, Bari, Italy, pp. 491-496, 2012.