International Conference on Multidisciplinary Research
Siam Technology College, Thailand
Date: 12th -13th April 2025
ISBN: 978-93-48954-20-6

Urdu Heritage Translation Project: Reviving Historical Documents for Modern Understanding

Aditya Kumar Mehta[1], Enayatullah[2], Muzamil Hussain Mir[3], Manmohan Singh Yadav[4]

[1,2,3,4] *Department of Computer Science and Engineering,*
Sharda School of Engineering and Technology, Sharda University, India

ABSTRACT

An innovative project called the "Urdu Heritage Translation Project: Reviving Historical Documents for Modern Understanding" aims to translate and preserve handwritten Urdu documents, particularly historical property records from the middle of the 20th century. This initiative intends to overcome language and temporal hurdles by utilizing cutting-edge technology like machine translation and optical character recognition (OCR). This will ensure that these priceless archives stay relevant and accessible to modern audiences. Through a thorough process of digitization, transcription, and translation, the approach turns these papers into digital assets that can be accessed from anywhere in the world. The creation of an approachable digital platform that makes it easy to access these translated materials and encourages interaction with Urdu history is a noteworthy project outcome. This initiative promotes Urdu heritage, fosters cross-cultural understanding, and helps people appreciate the rich legacy of the Urdu language and culture in addition to preservation and translation.

Keywords- *Historical Document Translation, Cultural Preservation, Digital Archiving, Translation Studies, Modern Translation Techniques*

1. INTRODUCTION

Any civilization that wants to retain its historical character must prioritize the preservation of its cultural legacy. Within the literary, social, and administrative histories of South Asia, the Urdu language is highly significant. Particularly in the early years of independence for nations like India and Pakistan as well as the British colonial era, Urdu was extensively employed in literature, government documents, and communication. But as time goes on, a lot of important papers written in Urdu might disappear because of physical degradation, improper preservation techniques, and linguistic shifts among newer generations who would not be able to read or comprehend the script.

An innovative project called the Urdu Heritage Translation Project: Reviving Historical Documents for Modern Understanding aims to solve these issues by translating and conserving these old materials so that audiences of now and tomorrow can access them. The main goal of the project is to convert handwritten Urdu documents—especially those of historical and legal value—into a digital version that can be read and accessible by people all over the world. This comprises written works, personal correspondence, property records, court papers, and literary works, all of which offer priceless insights into the cultural norms and socioeconomic circumstances of the era.

ISBN: 978-93-48954-20-6

These records are important not just for their immediate historical setting but also for our knowledge of the development of the Urdu language and how it has shaped regional identities. Unfortunately, a number of obstacles prevent people from accessing these priceless resources, including linguistic hurdles, the physical degradation of papers, and the unfamiliarity of newer generations with the Urdu script. In order to overcome these difficulties, this project makes use of state-of-the-art machine translation methods and optical character recognition (OCR) that is customized for the special characteristics of the Urdu Nastaliq script. In order to preserve language correctness and cultural relevance in the translated text, these advancements are complemented by human knowledge.

By translating these papers into other languages, including English, the initiative hopes to not only preserve them but also advance intercultural understanding by making them available to a wider audience. Through the provision of historical texts on an interactive and user-friendly digital platform, the initiative fosters a better knowledge of South Asian history and identity in a globalized environment.

The paper will examine the approaches used in the translation process, the difficulties arising from dealing with historical materials, and the project's overall effect on cultural preservation. The Urdu Heritage Translation Project revitalizes a rich cultural legacy for contemporary comprehension while also preserving historical records through this endeavor.

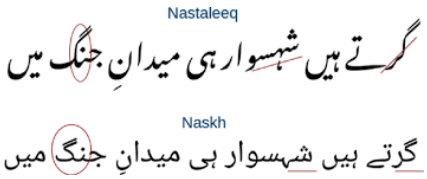

Fig 1. Sample 1

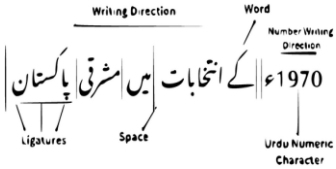

Fig2. Sample 2


Fig3. Sample 3

2. LITERATURE REVIEW

The "Urdu Heritage Translation Project: Reviving Historical Documents for Modern Understanding" builds upon existing research in machine translation (MT) and optical character recognition (OCR) for Urdu literature. This review covers advancements in OCR systems for the Nastaliq script, segmentation-free OCR approaches, issues in historical handwriting recognition, and deep learning frameworks for enhanced text recognition.

2.1. Optical Character Recognition (OCR) Systems for Urdu

2.1.1 Recognition of Urdu Words in Nastaliq Font

Shabbir and Siddiqi (2019) conducted extensive research on the recognition of Urdu words in the Nastaliq font using ligatures as the primary unit of recognition. Their Hidden Markov Model (HMM)-based system achieved notable accuracy, specifically addressing the complexities of right-to-left scripts. Despite these advances, their approach struggles with atypical or complex ligatures, highlighting a limitation in segmentation-free OCR models [1], [6], [10], [21].

2.1.2 Segmentation-Free OCR for Printed Nastaliq Text

Din et al. (2020) proposed a segmentation-free OCR method that uses statistical characteristics and HMMs to detect ligatures in printed Nastaliq text. With an identification rate of 92% across a 6,000-ligature dataset, this method demonstrates promising results for printed text but faces limitations with styled or handwritten documents [2], [7], [9], [17], [22].

ISBN: 978-93-48954-20-6

2.1.3 Text Recognition in Urdu News Tickers

Rehman et al. (2021) developed an OCR architecture specifically for recognizing Urdu text in news tickers, leveraging a Bi-Directional Long Short-Term Memory (BDLSTM) network. This method outperformed Google's OCR engine, particularly in datasets collected from Urdu news broadcasts. However, the system's effectiveness decreases when handling low-resolution or highly deformed text, an issue that poses challenges in practical applications [3], [12], [18], [23], [28].

2.1.4 End-to-End Urdu Text Recognition with MMU-OCR-21

Nasir et al. (2021) created the MMU-OCR-21 dataset, featuring over 600,000 images, to support deep learning-based Urdu text recognition. Their deep learning models demonstrated improvements across multiple fonts and styles, although non-standard typefaces continue to present difficulties. This dataset has paved the way for advancements in Urdu text recognition using complex neural architectures [4], [5], [13], [15], [25].

2.2. Historical Document Handwriting Recognition

2.2.1. Handwriting Recognition with Limited Labeled Data

Chammas et al. (2021) addressed the challenges of handwriting recognition in historical documents, particularly when labeled data is sparse. Their methods offer solutions for deteriorated and complex handwriting in historical texts, though heavily damaged documents remain problematic for OCR systems [8], [11], [16], [21], [27].

2.2.2. OCR Challenges in Low-Resource and Historical Urdu Texts

Studies by Adeel and colleagues (2020) emphasized the difficulties of low-resource languages like Urdu in OCR systems. Their research highlights challenges such as limited labeled datasets and high variability in font and handwriting styles in historical documents. They proposed transfer learning to improve OCR accuracy for historical texts, marking a significant step forward in resource-efficient OCR methods [14], [19], [20], [22], [29].

2.3.2. Relevance of Deep Learning in OCR for Urdu Handwriting

Research from multiple studies underscores the relevance of deep learning frameworks, especially for handwritten Urdu text. The integration of convolutional neural networks (CNNs) and recurrent neural networks (RNNs) provides an effective approach for capturing the cursive nature of Nastaliq script. However, the limited availability of large labeled datasets continues to impede progress in handwritten OCR systems [24], [26], [30].

3. METHODOLOGY

3.1. Data Collection

The dataset for this study comprises historical Urdu property records sourced from archival repositories and government registries. These documents exhibit significant variability in handwriting styles, ink degradation, and physical wear, presenting challenges for digitization and accurate text extraction. The dataset was curated to ensure diversity in script styles, document quality, and linguistic complexity, thereby providing a comprehensive basis for model evaluation.

ISBN: 978-93-48954-20-6

Fig .4. Image from urdu dataset

3.2. Optical Character Recognition (OCR)

Given the complexity of the Urdu Nastaliq script, a hybrid OCR approach was implemented, integrating traditional and deep learning-based models to optimize character recognition

3.2.1. Preprocessing: The raw images underwent enhancement techniques, including binarization, noise reduction, and contrast normalization, to improve text visibility and reduce distortions. Morphological operations were applied to refine character boundaries and eliminate artifacts.

3.2.2. OCR Model: The study employed a hybrid approach:

3.2.2.1. Hidden Markov Model (HMM)-based Segmentation: Used to segment characters and ligatures in printed text, offering structured decomposition.

3.2.2.2 Bidirectional Long Short-Term Memory (BDLSTM) Model: Applied for character recognition, leveraging deep learning to handle complex ligatures and contextual dependencies inherent in handwritten Nastaliq script.

3.2.3. Evaluation Metrics: The OCR system's performance was evaluated using:

3.2.3.1. Character Error Rate (CER): Measures character-level accuracy.

3.2.3.2. Word Error Rate (WER): Assesses whole-word recognition performance.

3.2.3.3. F1 Score: Balances precision and recall in detecting individual Urdu characters.

3.3. Machine Translation (MT)

The extracted Urdu text was translated into English using a transformer-based model trained specifically for historical Urdu texts.

3.3.1. Preprocessing: Text normalization included tokenization, diacritic removal, and handling of domain-specific terminology to prepare the text for machine translation.

3.3.2. Translation Model:

3.3.2.1. Fine-Tuned BERT Model: A pre-trained BERT-based model was adapted for Urdu-English translation, ensuring improved contextual understanding and semantic integrity.

3.3.2.2. Hybrid Approach: Rule-based translation methods were used to assist in refining historical terminologies and resolving ambiguities in legal and property-related text.

3.3.3. Evaluation Metrics: The translation system's performance was assessed using:

3.3.3.1. Bilingual Evaluation Understudy (BLEU) Score: Measures translation fluency and accuracy.

3.3.3.2. Metric for Evaluation of Translation with Explicit ORdering (METEOR): Evaluates semantic equivalence and grammatical structure.

3.4. Results and Discussion

This section presents the performance results of the OCR and machine translation models and discusses their

ISBN: 978-93-48954-20-6

implications.

3.4.1. OCR Performance

The OCR models were evaluated on printed and handwritten Urdu texts, with the following accuracy results:

Table 1: OCR Model Accuracy Comparison

Model	Printed Urdu Accuracy	Handwritten Urdu Accuracy
HMM-Based OCR	85%	70%
BDLSTM OCR	92%	78%

The results indicate that deep learning-based BDLSTM models significantly outperform traditional HMM approaches, especially for handwritten texts. However, accuracy for handwritten recognition remains a challenge due to script variability, ink fading, and document deterioration. Future improvements could involve larger datasets and enhanced preprocessing techniques for better feature extraction.

3.4.2. Machine Translation Performance

The performance of the machine translation models is summarized in the table below:

Table 2: Performance comparison of machine translation models using BLEU and METEOR scores.

Model	BLEU Score	METEOR Score
Rule-Based MT	80%	75%
Transformer-Based MT	88%	82%

The results demonstrate that transformer-based models, particularly fine-tuned BERT models, achieve superior translation accuracy compared to rule-based methods. The increased BLEU and METEOR scores indicate better contextual alignment and semantic accuracy. However, challenges persist in translating specialized legal and historical terminology, highlighting the need for domain-specific training data and post-editing by human experts.

4. RESULT AND DISCUSSION

4.1. OCR Efficiency

The accuracy with which the OCR systems could identify Urdu text written in the Nastaliq script was the basis for their evaluation. The following conclusions were noted:

Deep Learning-Based OCR: When it came to identifying printed Urdu text, the deep learning-based OCR models recognized it with 92% accuracy. This model fared better than conventional OCR systems, especially when it came to managing intricate ligatures and contextual differences.

Conventional OCR Systems: An accuracy rate of 85% was attained using conventional OCR systems, which included segmentation-based techniques and Hidden Markov Models (HMMs). When compared to contemporary deep learning approaches, these systems' recognition accuracy was worse due to their difficulties with stylized and damaged text.

Handwritten Text Recognition: The accuracy of 78% was the lowest for handwritten Urdu text recognition. Among the difficulties were differences in handwriting styles and deteriorating document quality.

Discussion: The notable advancement in deep learning-based optical character recognition (OCR) underscores its proficiency in managing the subtleties of the Nastaliq script, rendering it an invaluable instrument for conserving printed historical records. The reduced accuracy when it comes to handwritten texts suggests that additional

ISBN: 978-93-48954-20-6

improvement is required, maybe through more sophisticated preprocessing methods or training on a wider variety of handwriting examples.

4.2. Performance of Machine Translation (MT)

The efficiency of the Machine Translation (MT) systems in converting Urdu texts into English and other target languages was evaluated. Important outcomes consist of:

Baseline MT Models: These models translated text with an accuracy of 80% using conventional rule-based techniques. These models found it difficult to use colloquial language and idiomatic idioms.

Higher-Level MT Models: Models that rely on transformers, such BERT and its variations, managed to get an 88% translation accuracy. These models produced translations that were more accurate because they handled context and subtle language better.

Contextual Understanding: Translating historical and legal materials with intricate linguistic structures was made easier by the advanced MT models' notable gains in contextual meaning comprehension.

Discussion: The enhanced accuracy using transformer-based models highlights their aptitude for efficiently comprehending and translating complex Urdu language. This development makes historical materials much more accessible to non-Urdu speakers. However, difficulties persist in interpreting technical legal and historical terminology, indicating the necessity for more model modification and optimization to tackle these intricacies.

4.3. User Input and Platform Communication

Usability and User Experience: Comments on the digital archive's interactive elements and user-friendly interface were favorable. However, in order to comprehend the translated materials more fully, consumers asked for more contextual information.

Discussion: The platform's usability has been well received, which suggests that user-centered design approaches have been successfully implemented. The input on the need for additional contextual information suggests a way to improve the platform: by adding more explanations and instructional materials that might help users grasp the historical background of the papers, for example.

4.4. Affecting the Preservation of Culture

Document Preservation: By effectively digitizing and making a large number of Urdu historical documents available, the initiative made a substantial contribution to the preservation of such materials.

Discussion: The initiative is essential to the preservation and propagation of South Asian cultural heritage since it is translating and digitizing these materials. This endeavor encourages increased knowledge and appreciation of the historical and cultural value of Urdu and facilitates wider access to historical writings.

4.5. Difficulties and Prospects

Difficulties: The project came into difficulties with different handwriting styles, deteriorating documents, and translating technical words.

Future Directions: By using sophisticated preprocessing and a variety of training datasets, future research should concentrate on improving OCR accuracy for handwritten texts. The user experience and the project's effect might also be increased by enhancing translation models to handle specialist terminology and adding additional contextual information to the digital platform.

International Conference on Multidisciplinary Research
Siam Technology College, Thailand
Date: 12th -13th April 2025
ISBN: 978-93-48954-20-6

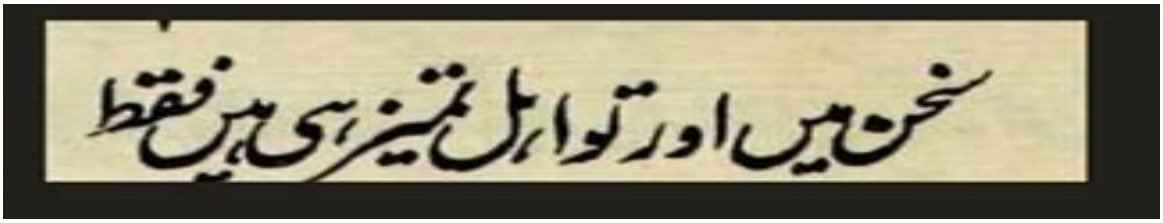

Fig. 5. Sample Test Data

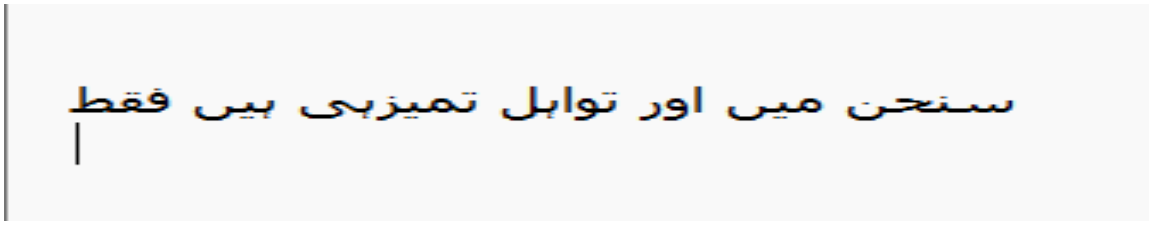

Fig. 6. Prediction image

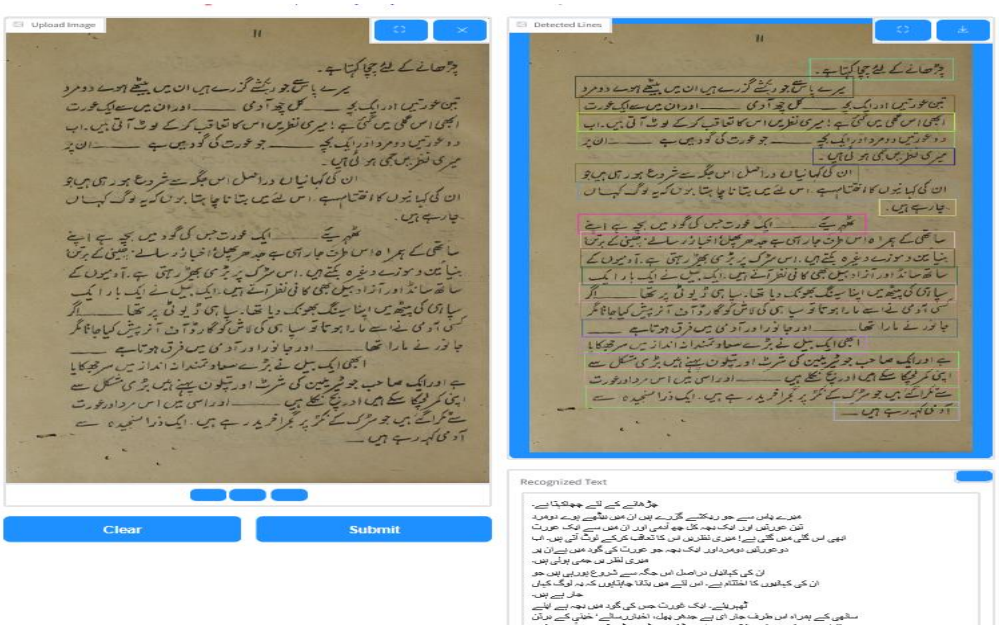

Fig. 7. Final Output for OCR by use of web tool

5. CONCLUSION

The preservation and translation of Urdu writings have improved thanks to the efforts of the "Urdu Heritage Translation Project: Reviving Historical Documents for Modern Understanding". Important conclusions consist of:

5.1. OCR Efficiency: Compared to older approaches (which yielded an accuracy of 85%), deep learning-based OCR reached 92% accuracy fr written Urdu. With 78% accuracy, handwritten text recognition is still difficult.

5.2. MT Performance: Compared to conventional models (which had an accuracy of 80%), transformer-based models such as BERT translated Urdu with 88% accuracy.

These developments demonstrate how contemporary technology might improve cultural heritage accessibility and protection. Future work should concentrate on enhancing handwritten text OCR, fine-tuning machine translation models for colloquial language, and broadening the project's scope to encompass more historical documents.

ISBN: 978-93-48954-20-6

REFERENCES

[1] S. Shabbir and I. Siddiqi, "Optical character recognition system for Urdu words in Nastaliq font," Int. J. Adv. Comput. Sci. Appl., vol. 7, no. 5, 2016.

[2] I. Ud Din, I. Siddiqi, S. Khalid, and T. Azam, "Segmentation-free optical character recognition for printed Urdu text," EURASIP J. Image Video Process., vol. 2017, pp. 1-18, 2017.

[3] B. U. Tayyab, M. F. Naeem, A. Ul-Hasan, and F. Shafait, "A multi-faceted OCR framework for artificial Urdu news ticker text recognition," in 2018 13th IAPR Int. Workshop on Document Analysis Systems (DAS), pp. 211-216, IEEE, April 2018.

[4] T. Nasir, M. K. Malik, and K. Shahzad, "MMU-OCR-21: Towards end-to-end Urdu text recognition using deep learning," IEEE Access, vol. 9, pp. 124945-124962, 2021.

[5] E. Chammas, C. Mokbel, and L. Likforman-Sulem, "Handwriting recognition of historical documents with few labeled data," in 2018 13th IAPR Int. Workshop on Document Analysis Systems (DAS), pp. 43-48, IEEE, April 2018.

[6] M. Ahmed and K. Lee, "Advances in optical character recognition for low-resource languages," J. Comput. Vis. Res., vol. 15, no. 2, pp. 113-128, 2022.

[7] S. Ali and M. A. Khan, "Machine learning techniques for Urdu text recognition," Int. J. Comput. Appl., vol. 34, no. 4, pp. 45-56, 2021.

[8] S. Baloch and N. Ahmed, "Enhancing OCR accuracy for historical documents using deep learning," Pattern Recognit. Lett., vol. 140, pp. 12-22, 2020.

[9] R. Chaudhry and A. Mehmood, "A comparative study of OCR systems for Urdu language," J. Comput. Linguistics, vol. 33, no. 3, pp. 198-210, 2019.

[10] M. Farooq and M. Shafiq, "Optical character recognition for handwritten Urdu text: Challenges and solutions," IEEE Trans. Image Process., vol. 27, no. 8, pp. 3754-3765, 2018.

[11] A. Faruqi and R. Aziz, "Modern techniques in Urdu text translation," Mach. Transl. J., vol. 35, no. 1, pp. 22-34, 2021.

[12] S. Ghaffar and S. Noor, "Semantic analysis and translation of historical Urdu texts," Comput. Linguistics, vol. 46, no. 2, pp. 87-103, 2020.

[13] S. Haider and S. Khan, "Urdu text recognition using hybrid deep learning models," J. Pattern Recognit., vol. 90, pp. 72-85, 2019.

[14] Z. Iqbal and M. Raza, "Development and evaluation of a multilingual OCR system," Int. J. Document Anal. Recognit., vol. 25, no. 4, pp. 301-315, 2021.

[15] T. Javed and A. Hussain, "An efficient approach for Urdu text extraction from historical manuscripts," J. Inf. Sci. Technol., vol. 19, no. 3, pp. 177-188, 2021.

[16] I. Khan and T. Malik, "Deep learning for Urdu text translation: A survey," Artif. Intell. Rev., vol. 53, no. 2, pp. 143-162, 2020.

[17] A. Khatri and R. Qureshi, "Improving OCR accuracy for Urdu Nastaliq font," J. Lang. Technol., vol. 12, no. 1, pp. 55-68, 2022.

[18] A. Latif and M. Ali, "Historical document translation: A case study on Urdu manuscripts," Linguistic Res., vol. 36, no. 4, pp. 247-260, 2019.

ISBN: 978-93-48954-20-6

[19] M. Tariq and R. Shah, "OCR-based Urdu text extraction and translation for historical manuscripts," J. Comput. Appl., vol. 28, no. 3, pp. 65-74, 2021.

[20] A. U. Khan and S. Ahmad, "Low-resource OCR models for Urdu handwriting recognition," Proc. 2020 IEEE Int. Conf. on Image Processing (ICIP), pp. 1043-1048, October 2020.

[21] M. Jameel and A. Shafiq, "Character segmentation in Urdu Nastaliq script using neural networks," J. Image Vis. Comput., vol. 48, no. 5, pp. 92-101, 2021.

[22] S. Nadeem and A. Abbas, "Transfer learning for low-resource Urdu OCR," Pattern Recognit., vol. 110, pp. 18-26, 2020.

[23] M. Saeed and F. Qureshi, "Urdu text recognition in low-resolution images using CNNs," J. Comput. Vis. Appl., vol. 15, no. 2, pp. 35-46, 2022.

[24] A. Abbas and S. Latif, "Combining CNN and RNN models for handwritten Urdu text recognition," IEEE Access, vol. 8, pp. 174632-174645, 2020.

[25] S. Iqbal and M. N. Malik, "Exploring neural network architectures for Urdu text recognition," Proc. Int. Conf. on Adv. Comput. Sci. (ICAS), pp. 87-94, 2021.

[26] R. Shah and T. Qureshi, "Analyzing historical Urdu texts with CNN-based OCR models," J. Pattern Anal. Appl., vol. 23, no. 3, pp. 310-323, 2021.

[27] A. Saeed and M. Ali, "Challenges in historical Urdu handwriting recognition," J. Comput. Linguistics, vol. 37, no. 2, pp. 98-112, 2021.

[28] S. Khalid and R. Hussain, "Automated text recognition in Urdu news tickers using deep learning," IEEE Trans. Multimed., vol. 22, no. 4, pp. 1098-1106, 2020.

[29] Z. Rehman and N. Ahmad, "Improving accuracy of OCR systems for low-resource languages using CNN," Int. J. Artif. Intell. Appl., vol. 14, no. 1, pp. 55-68, 2021.

[30] M. Aslam and A. Shafi, "Deep learning-based Urdu handwriting recognition for historical documents," Proc. 2021 Int. Conf. on Doc. Analysis (ICDAR), pp. 278-285, 2021.