Natural Resources for Human Health
eISSN: 2583-1194

Original Research

Received 15 May 2026
Revised 18 June 2026
Accepted 08 July 2026

Natural Resources for Human Health 2026, Vol. 6 (6s): 1322–1332

KEYWORDS:
Online handwriting recognition, bilingual script, text recognition, devanagari, roman, handwritten text.

An Efficient and Scalable Approach for Script Recognition in Online Bilingual Handwritten Text

Mamta[1]*, Punit Soni[1], Gurpreet Singh[2]

[1]Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India
[2]School of Computer Science and Engineering, Lovely Professional University, Punjab, India

ABSTRACT: Script identification of handwritten text is one of the most captivating and complex applications for recognizing of different patterns of text in the field of pattern recognition. There's been a lot of progress in recognizing handwriting in monolingual environment, very few have explored in bilingual or mixed-script environment. In this paper, a Handwriting Recognition System is proposed for the code-mixed text for bilingual words of two different scripts (Devanagari for Hindi Language and Roman for English language). To recognize the script of a segmented handwritten word, different classification algorithms have been used and then comparative analysis has been performed to measure the recognition performance rate of MLP, SVM, HMM and DNM (Deep Net Model). First pass of the proposed system identified or segmented the words based on their connectivity. After that these segmented words, further segmented at the level of stroke points. From the identified types of these stroke points the idea about the script associated with the complete word has been recognized. In this proposed system, 97.025% accuracy have been achieved as per stroke identification process and 98.07% with script identification. After script identification, on the basis of these results an appropriate handwriting recognition engine can be applied to the input to convert it to its digital form.

1. INTRODUCTION

In many multilingual societies, people often mix languages in both spoken and written communication a practice known as code-mixing [1-2]. This is especially common in places like India, where languages such as Hindi, English, Tamil, and others are frequently combined in daily conversations and informal written forms. As more people use touchscreen devices to write and share handwritten text online, code-mixed handwriting presents new challenges for handwriting recognition system [3-4]. Script recognition is an essential step in the process of recognizing this type of handwritten text. It involves in determining which handwriting recognition engine is required to convert the handwritten inputs to its digital form [5-6]. The applications of automatic handwriting recognition in the computer system are labelled as online systems or offline systems. In case of online systems, the input is given with the help of digitizer or active stylus, whereas in case of offline systems, scanned image containing handwritten text is taken as input [7-10]. The pre-processing of stroke input, segmentation, feature extraction, and classification have been determined to represent the key operations in these OHR systems [11]. Every monolingual OHR system is created by taking into account the script's challenges and limitations. However, recognizing the script of each handwritten word before transforming it to its digital format is an initial obstacle in bilingual systems. The varied character sets and different features of each script make this stage extremely hard and complicated [12-15].

2. LITERATURE REVIEW

Many researchers proposed system for the recognition of bilingual scripts, such as- Sheth et al. [16] presented COMI-LINGUA, the largest manually annotated Hindi-English code-mixed dataset with over 125K instances covering five NLP tasks: matrix language identification, token-level identification, POS tagging, NER, and machine translation. Ojo et al. [17] presented a CK-Keras model with pre-trained Word2Vec embedding for achieving better results. Author focused on identifying languages at the word level in Kannada-English (Kn-En) code-mixed YouTube comments. Banerjee et al. [18] focused on extracting named entities (NE) from code-mixed, cross-script social media data. A Bengali-English (Bn-En) dataset was introduced along with domain-specific NE taxonomies. Using both formal and informal language features, four machine learning algorithms—CRF, MIRA, SVM, and MEMM—were applied for NE recognition. Bengali was treated as the native language and English as the non-native. Among the models, CRF and SVM performed best, and the approach can generalize to other code-mixed datasets. Mamta et al. [19] proposed a system for word segmentation technique for code-mixed handwritten text. It explores a bilingual OHR system for code-mixed text in Roman (English) and Devanagari (Hindi). Their proposed system achieved an accuracy of 92.975%. Constum.T [20] presented DANIEL (Document Attention Network for Information Extraction and Labelling)—a fully end-to-end system that combines layout recognition, handwriting recognition, and entity extraction in one architecture. Using a convolutional encoder and a transformer-based decoder, DANIEL works across languages, layouts, and tasks. Singh et al. [21] explored the growing field of code-mixing, a natural outcome of globalization and diverse communication platforms like social media, messaging apps, and online forums. It synthesizes research from linguistic, sociolinguistic, and computational viewpoints, offering a broad overview of key themes, methods, and findings. As this is a literature review, it does not report any specific accuracy. but rather highlights prior research contributions. Chaudhary et al. [22] suggested a pattern analysis approach for a monolingual system. The process starts with tokenizing words, identifying their language using confidence scores, and applying a threshold to determine dominant languages. Language rules are used to identify script while fallback mechanisms handle ambiguities. Translation tools or language-specific libraries are then used to create a unified output.

3. PROPOSED MODEL

This paper presents a technique of script identification of bilingual online handwriting system. The handwritten text supplied into the system contains a code-mixed words written using Hindi and English words. Devanagari script is used to write Hindi words and Roman script is used to write English words. Each script has different character sets [23,24]. Figure-1 represents the overall structure of the proposed system. Bilingual code-mixed input is supplied to the system by using digitizer and stylus in the form of online handwriting strokes [25,26]. After taking the input, pre-processing steps have been performed to deal with the existence of similar strokes points, identification of missing stroke points and stroke normalization. SSPR, LMPV and NCS algorithms respectively have been proposed to deal with pre-processing stage. After preforming pre-processing, the identification of connected strokes has been performed for the segmentation at the level of different words present in the sample input. This word level segmentation is followed by stroke identification for respective words [25-28]. Four different algorithms MLP, SVM, HMM and DNM (proposed model) have been performed for the identification of different strokes involved in the formation of a complete word. The proposed model DNM performed exceptionally well in comparison with the other models such as MLP, SVM and HMM. A total of 91 unique strokes (presented in Table-1) has been identified to deal with the issue of handwriting recognition in case of code mixed (Hindi-English). Table-1 shows the selected words for the collection of handwriting samples for dataset. The base criteria for the selection of only these words are the appearance of all the characters from the character set of Devanagari script.

Natural Resources for Human Health
eISSN: 2583-1194

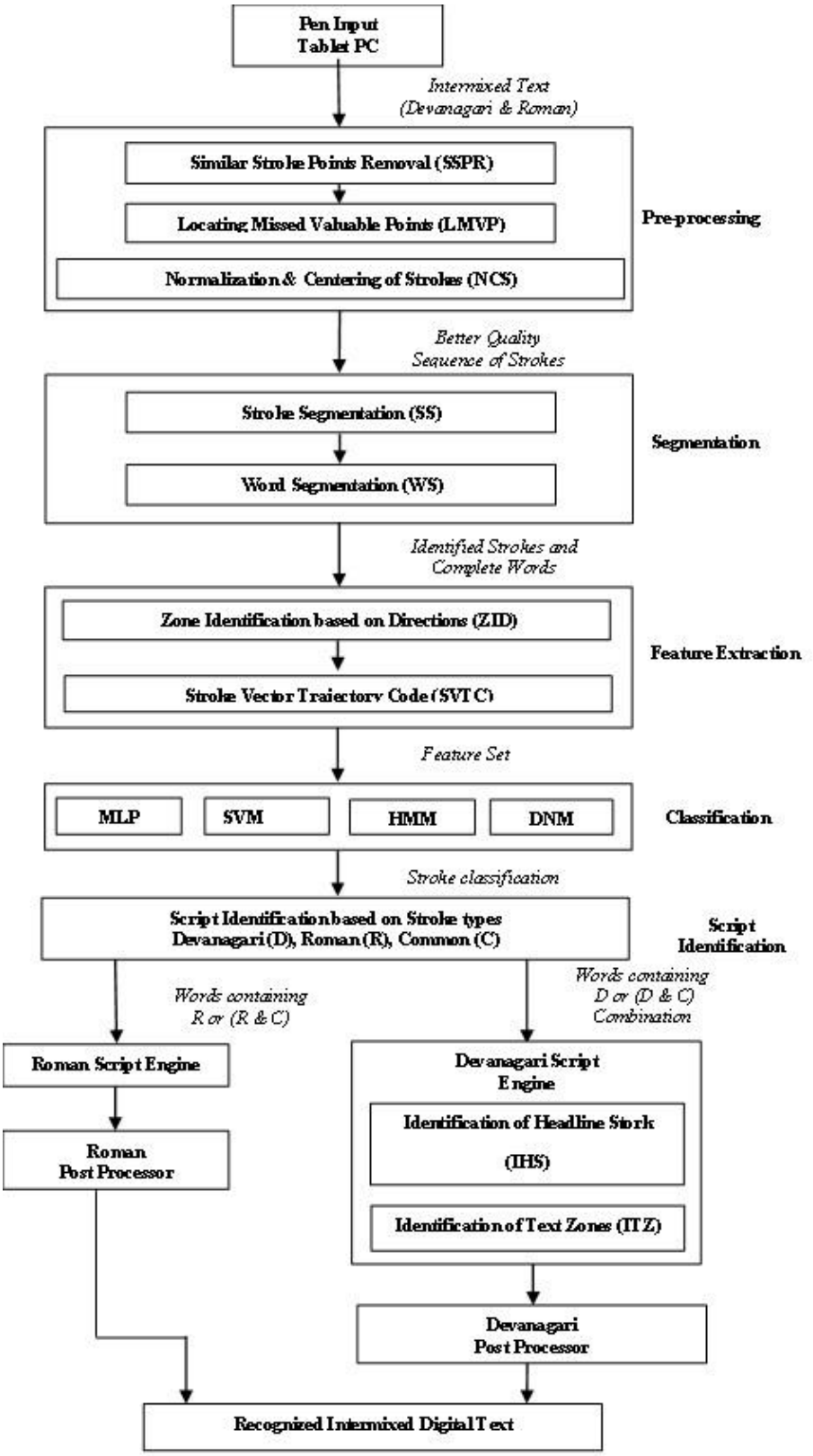


Fig 1: Framework of Online Bilingual (Devanagari-Roman Text)

Natural Resources for Human Health
eISSN: 2583-1194

Table 1. Selected words for dataset collection

अमृतसर	चितौड.गढ.	पंढरपुर
आन्ध्रप्रदेश	चित्रदुर्ग	पीलीभीत
अराकोणम	छिंदवाड.ा	पोर्टब्लेयर
अल्मोड.ा	जोशीमठ	बदायुं
उधमपुर	झुमरी तिलैया	भुवनेश्वर
उत्तरकाशी	टांडा	मथुरा
ऋषिकेश	डाल्टेनगंज	मधुबनी
औराई	ढेंकानाल	रणथम्भोर
कडलूर	तिरूवनन्तपुरम	लखनऊ
कच्छ	तोशाम	श्रीकाकुलम
काशीपुर	ठाणे	संगरूर
खजुराहो	दार्जिलिंग	हल्द्वानी
गाँधीनगर	धंधुका	कुरुक्षेत्र
गुलबर्ग	नजीबाबाद	ज्ञानयापी
ग्रेटरनोएडा	नंदप्रयाग	
गढ.वा	नांदेड.	
गोण्डा	नजफगढ.	
घरौंडा	नौगछिया	

These strokes further categorized under three different categories as Devanagari (D) strokes, Roman (R) strokes and Common (C) strokes. As per the findings of this study, it has been observed sometimes the mixture of Roman-Common or Devanagari-Common [29,30]. This further helps in the identification of the script associated with the segmented word under consideration. The capturing of the directional movement of stylus with the help of ZID (Zone identification based on direction) and SVTC (Stroke Vector Trajectory Code) algorithms helps to identify the stroke and its category. The application of classification algorithms further helps to identify the strokes and their categories. Figure-2a,2b and 2c represented the handwritten input, word segmentation and pre-processed outputs respectively for the sample code mixed text.

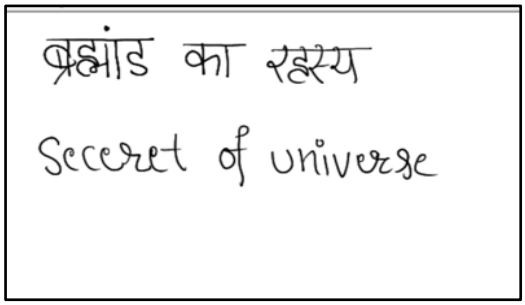


Fig 2a. Snapshot of the input given to the system

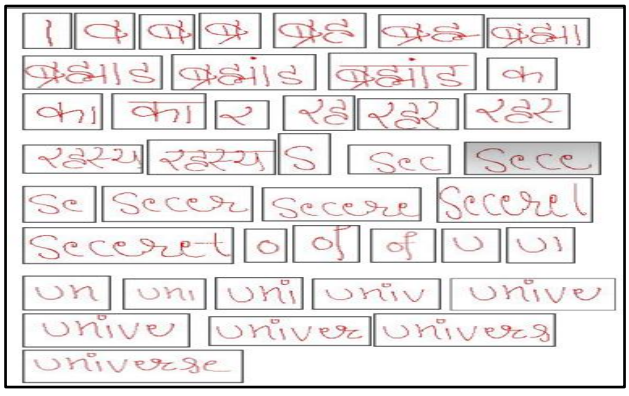


Fig 2b. Extracted strokes from the handwritten input

Natural Resources for Human Health
eISSN: 2583-1194

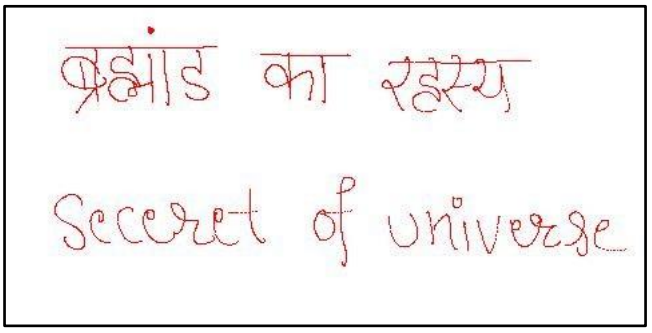


Fig 2c. pre-processed form of handwritten input

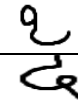
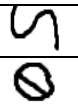
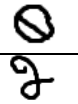
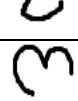
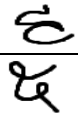
4. RESULT AND DISCUSSION

The experimental results of the proposed approach for the handwritten Devanagari–Roman bilingual script identification in online handwriting recognition are presented and discussed in this section. The effectiveness and robustness of the proposed model are assessed by applying the standard classification metrics and compared with the existing baseline models.

Table-2 show the 91 uniquely identified strokes. Out of these 91 strokes classes, 12 stroke classes have been considered under the category of Common (C) strokes. These common strokes can be used to write the characters of Devanagari as well as Roman scripts. 37 stroke categories reflect Devanagari (D) type of strokes and 42 stroke categories belongs to Roman (R) strokes. Figure-3 shows the comparison of MLP, SVM, HMM and proposed model for stroke identification process. This further helps in the identification of the script related to the segmented word. After identification of the script the respective handwriting recognition system will help to generate the digital output with respect to the segmented handwritten word.

Table 2: Stroke Set Recognition Accuracy

S. No	Stroke	Stroke Type	MLP	SVM	HMM	DNM
1		D	87.49	82.39	86.46	95.87
2		D	99.01	94.26	98.33	100
3		D	94.64	95.22	99.56	100
4		D	91.97	91.23	95.8	98.73
5		D	83.12	84.48	87.09	94.82
6		D	83.12	86.45	89.01	95.68
7		D	81.16	85.27	89.28	96.64
8		D	97.32	88.55	91.84	98.67
9		D	92.02	80.51	84.05	94.82
10		D	94.16	82.16	86.16	97.98
11		D	80.41	80.63	85.09	93.67

12		D	99.4	92.73	94.48	100
13		D	96.65	86.29	88.23	99.95
14		D	84.25	87.45	88.01	95.76
15		D	83.64	85.4	88.33	96.78
16		D	83.67	84.99	85.26	93.56
17		D	86.08	88.21	90.59	96.83
18		D	90.5	94.99	97.75	98.96
19		D	88.64	84.58	86.08	93.77
20		D	85.82	88.54	84.53	92.68
21		D	92.24	85.8	86.04	97.72
22		D	82.79	83.22	83.51	93.64
23		D	85.84	86.73	90.87	94.83
24		D	87.33	91.83	93.7	98.81
25		D	89.12	92.19	92.92	96.72
26		D	95.7	95.85	98.51	100
27		D	83.99	84.59	88.46	94.95
28		D	90.28	83.73	84.89	96.73
29		D	91.85	91.97	95.13	99.96
30		D	80.93	81.82	82.34	93.89
31		D	92.15	94.94	95.29	100
32		D	83.41	86.9	89.6	94.73
33		D	81.3	84.6	87.34	92.89
34		D	98.98	82.2	85.42	99.62
35		D	99.31	84.56	88.22	100
36		D	96.17	88.54	93.42	98.81
37		D	86.09	87.79	90.42	94.72
38		D	91.95	85.71	87.39	93.87

39		R	93.68	80.14	84.14	96.91
40		R	88.8	90.21	91.64	95.84
41		R	82.44	85.76	88.01	93.82
42		R	89.9	84.44	84.93	96.74
43		R	80.69	81.25	81.47	94.91
44		R	98.19	86.75	91.57	100
45		R	85.18	86.58	90.77	93.84
46		R	93.25	86.46	89.97	96.67
47		R	86.23	90.38	92.48	94.81
48		R	90.4	92.43	93.38	95.73
49		R	90.93	87.27	88.14	96.95
50		R	83.7	86.89	88.22	94.92
51		R	99.39	99.25	100	100
52		R	95.5	85.04	88.64	97.83
53		R	98.79	89.94	93.28	99.96
54		R	97.9	86.02	87.49	98.76
55		R	91.96	85.7	90.48	96.84
56		R	98.44	80.74	84.45	100
57		R	81.77	83.25	86.06	93.65
58		R	83.92	84.14	87.24	95.85
59		R	89.97	81.03	83.19	96.82
60		R	86.51	85.57	86.89	94.97
61		R	87.77	92.48	94.33	98.75
62		R	85.43	84.79	88.6	94.76
63		R	96.57	82.9	83.07	99.78

64		R	87.14	89.05	89.72	97.69
65		R	85.62	85.79	86.12	95.84
66		R	90.85	84.84	85.14	96.93
67		R	82.82	85.02	89.31	93.94
68		R	96.04	95.23	98.78	100
69		R	81.49	84.75	87.18	95.89
70		R	99.74	94.56	95.14	100
71		R	95.44	87.36	89.86	97.87
72		R	83.97	85.96	88.38	99.81
73		R	88.11	84.38	85.33	94.86
74		R	96.31	90.72	92.94	99.87
75		R	94.14	81.81	83.86	98.92
76		R	94.58	96.71	99.82	100
77		R	95.43	86.42	89.63	98.83
78		R	81.48	83.73	84.05	94.79
79		R	87.17	82.82	82.75	95.83
80		C	82.32	82.89	86.06	93.99
81		C	97.26	93.55	96.12	99.89
82		C	92.47	80.33	84.63	96.93
83		C	91.62	87.4	90.73	95.97
84		C	91.27	83.91	84.81	96.91
85		C	86.22	90.62	91.06	97.86
86		C	86.5	83.49	86.74	96.84
87		C	94.59	93.82	94.05	99.91
88		C	92.75	87.73	90.71	97.93

89		C	97.74	98.73	100	100
90		C	89.44	82.75	85.67	95.89
91		C	92.15	94.94	98.29	100

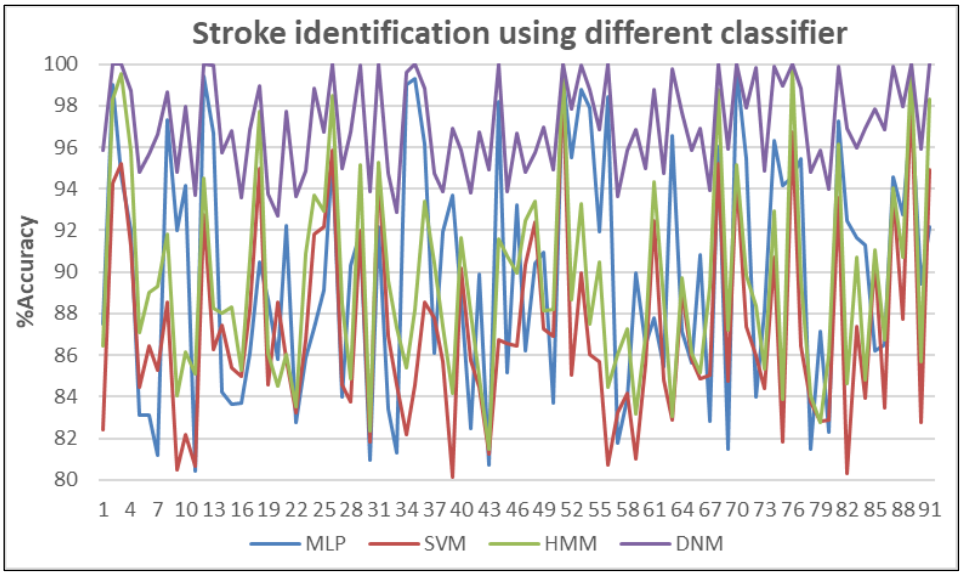


Fig 3: Stroke Identification using Classification Technique

5. CONCLUSION

The work presented here provides an effective and scalable solution for online bilingual handwritten text script identification. The scripts considered in this work are Devanagari and Roman. The suggested script identification algorithm extracts structural, directional, and dynamic characteristics from the input strokes, and classifies them with different classifiers MLP, SVM, HMM and DNM (Proposed model). The framework is able to correctly determine if a handwritten character is written in Hindi or English, building upon the feature-classifier pipeline. The accuracies achieved through the DNM model, demonstrated how robust the model remained to variations in writing style, speed, and structure while being computationally efficient as compared to traditional models. The proposed DNM model has achieved the 97.02% accuracy for stroke identification process and 98.07% accuracy for script identification process. The framework's scalability allows for variations to be made to support any other scripts, and could be used for multilingual handwriting recognition solutions for real-world applications such as digital archiving, education, and document processing. Hybrid deep neural network and transfer learning model can be considered for the future work for large dataset or multilingual framework.

ACKNOWLEDGEMENTS

The authors would like to thank everyone who took part in this data collection process and contributed to this research. The authors would also like to express their gratitude to the staff and facilities of their respective institutions which helped them conduct the study successfully.

AUTHOR CONTRIBUTIONS

Mamta: Conceptualization, Investigation, Writing and Editing.

Punit Soni: Methodology, Analysis and Supervision

Gurpreet Singh: Supervision, Writing—Review & Editing

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

ETHICS STATEMENT

Ethical approval was not required for this study.

DATA AVAILABILITY STATEMENT

Data supporting the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

[1] Sristy, N.B., Krishna, N.S., Krishna, B.S. and Ravi, V., 2017. Language identification in mixed script. In: Proceedings of the 9th Annual Meeting of the Forum for Information Retrieval Evaluation, pp. 14–20.

[2] Hidayatullah, A.F., Qazi, A., Lai, D.T.C. and Apong, R.A., 2022. A systematic review on language identification of code-mixed text: techniques, data availability, challenges, and framework development. IEEE Access, 10, pp. 122812–122831.

[3] Kazi, M., Mehta, H. and Bharti, S., 2020. Sentence level language identification in Gujarati-Hindi code-mixed scripts. In: 2020 IEEE International Symposium on Sustainable Energy, Signal Processing and Cyber Security (iSSSC), pp. 1–6.

[4] Lamabam, P. and Chakma, K., 2016. A language identification system for code-mixed English-Manipuri social media text. In: 2016 IEEE International Conference on Engineering and Technology (ICETECH), pp. 79–83.

[5] Dey, S., Thakur, S., Kandwal, A., Kumar, R., Dasgupta, S. and Roy, P.P., 2024. BharatBhasaNet: A unified framework to identify Indian code-mix languages. IEEE Access, 12, pp. 68893–68904.

[6] Francis, S.A. and Sangeetha, M., 2025. A comparison study on optical character recognition models in mathematical equations and in any language. Results in Control and Optimization, 18, p. 100532.

[7] Singh, G. and Sachan, K., 2019. A bilingual (Gurmukhi-Roman) online handwriting identification and recognition system. International Journal of Recent Technology and Engineering, 8(1), pp. 2936–2952.

[8] Valikhani, S., Abdali-Mohammadi, F. and Fathi, A., 2020. Online continuous multi-stroke Persian/Arabic character recognition by novel spatio-temporal features for digitizer pen devices. Neural Computing and Applications, 32, pp. 3853–3872.

[9] Wehbi, M., Hamann, T., Barth, J. and Eskofier, B., 2020. Digitizing handwriting with a sensor pen: A writer-independent recognizer. In: 2020 17th International Conference on Frontiers in Handwriting Recognition (ICFHR), pp. 295–300.

[10] Alhamad, H.A., Shehab, M., Shambour, M.K.Y., Abu-Hashem, M.A., Abuthawabeh, A., Al-Aqrabi, H. et al., 2024. Handwritten recognition techniques: A comprehensive review. Symmetry, 16(6), p. 681.

[11] Singh, H., Sharma, R.K. and Singh, V.P., 2021. Online handwriting recognition systems for Indic and non-Indic scripts: A review. Artificial Intelligence Review, 54(2), pp. 1525–1579.

[12] Mamta and Singh, G., 2023. An automatic process of online handwriting recognition and its challenges. In: International Conference on Deep Learning, Artificial Intelligence and Robotics, pp. 387–394. Cham: Springer Nature Switzerland.

[13] Tasdemir, E.F.B. and Yanikoglu, B., 2019. A comparative study of delayed stroke handling approaches in online handwriting. International Journal on Document Analysis and Recognition (IJDAR), 22, pp. 15–28.

[14] Manimaran, A., Syed, M.H., Kumar, M.S., Selvanayaki, S., Sunitha, G. and Manna, A., 2024. Enhancing Asian Indigenous Language Processing through deep learning-based handwriting recognition and optimization techniques. ACM Transactions on Asian and Low-Resource Language Information Processing, 23(8), pp. 1–20.

[15] Sheth, R., Beniwal, H. and Singh, M., 2025. COMI-LINGUA: Expert annotated large-scale dataset for multitask NLP in Hindi-English code-mixing. arXiv preprint, arXiv:2503.21670.

[16] Ojo, O.E., Gelbukh, A., Calvo, H., Feldman, A., Adebanji, O.O. and Armenta-Segura, J., 2022. Language identification at the word level in code-mixed texts using character sequence and word embedding. In: Proceedings of the 19th International Conference on Natural Language Processing (ICON): Shared Task on Word Level Language Identification in Code-mixed Kannada-English Texts, pp. 1–6.

[17] Banerjee, S., Naskar, S.K., Rosso, P. and Bandyopadhyay, S., 2017. Named entity recognition on code-mixed cross-script social media content. Computación y Sistemas, 21(4), pp. 681–692.

[18] Singh, G., 2024. A word segmentation approach for code-mixed handwritten text. In: 2024 4th International Conference on Sustainable Expert Systems (ICSES), pp. 526–532.

[19] Constum, T., Tranouez, P. and Paquet, T., 2025. DANIEL: A fast Document Attention Network for information extraction and labelling of handwritten documents. International Journal on Document Analysis and Recognition (IJDAR), pp. 1–23.
[20] Singh, S.P. and Mangla, N., 2024. Code-mixed text: An extensive literature review. In: 2024 Sixth International Conference on Computational Intelligence and Communication Technologies (CCICT), pp. 417–422.
[21] Chaudhary, A. and Shekhar, S., 2024. A framework to find single language version using pattern analysis in mixed script queries. In: 2024 2nd International Conference on Disruptive Technologies (ICDT), pp. 1593–1601.
[22] Singh, G. and Soni, P., 2024. Performance comparison of machine learning algorithms for online handwriting recognition: A brief analysis. In: 2024 4th International Conference on Sustainable Expert Systems (ICSES), pp. 1464–1469.
[23] Ghosh, T., Sen, S., Obaidullah, S.M., Santosh, K.C., Roy, K. and Pal, U., 2022. Advances in online handwritten recognition in the last decades. Computer Science Review, 46, p. 100515.
[24] Singh, G. and Sachan, M.K., 2020. An unconstrained and effective approach of script identification for online bilingual handwritten text. National Academy Science Letters, 43(5), pp. 453–456.
[25] Vinjit, B.M., Bhojak, M.K., Kumar, S. and Chalak, G., 2020. A review on handwritten character recognition methods and techniques. In: 2020 International Conference on Communication and Signal Processing (ICCSP), pp. 1224–1228.
[26] Keysers, D., Deselaers, T., Rowley, H.A., Wang, L.L. and Carbune, V., 2016. Multi-language online handwriting recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(6), pp. 1180–1194.
[27] ul Sehr Zia, N., Naeem, M.F., Raza, S.M.K., Khan, M.M., Ul-Hasan, A. and Shafait, F., 2022. A convolutional recursive deep architecture for unconstrained Urdu handwriting recognition. Neural Computing and Applications, 34(2), pp. 1635–1648.
[28] Chakraborty, B., Mukherjee, P.S. and Bhattacharya, U., 2016. Bangla online handwriting recognition using recurrent neural network architecture. In: Proceedings of the Tenth Indian Conference on Computer Vision, Graphics and Image Processing, pp. 1–8.
[29] Ismail, S.M., Abdullah, S.N.H.S. and Norul, S., 2012. Online Arabic handwritten character recognition based on a rule-based approach. Journal of Computer Science, 8(11), pp. 1859–1868.