International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Advanced Scene Text and Handwriting Recognition for Hindi Using Synthetic Data and Transfer Learning

Barkha Sahu

Submitted: 15/11/2022 Revised: 05/01/2023 Accepted: 14/01/2023

Abstract*:* The recognition of scene text and handwritten characters in the Hindi language presents significant challenges due to the complexity of the Devanagari script, diverse font styles, and limited annotated datasets. This paper proposes an advanced framework for **Scene Text Recognition (STR)** and **Handwriting Recognition (HWR)** in Hindi by leveraging **synthetic data generation** and **transfer learning** methodologies. Synthetic datasets enriched with both Unicode and non-Unicode fonts, along with varied handwritten styles, were developed to address data scarcity. Transfer learning techniques, adapted from pre-trained models on extensive multilingual datasets, significantly improved recognition performance by enabling cross-script knowledge transfer. Experimental results demonstrated a **33% improvement in Word Recognition Rate (WRR)** on the IIIT-ILST Hindi dataset, validating the effectiveness of our approach. Additionally, transfer learning across six Indian languages revealed potential inter-script benefits, where Hindi models benefited more from Indian scripts than from English datasets. Our work highlights the importance of synthetic data augmentation and cross-lingual learning for enhancing the accuracy and robustness of Hindi STR and HWR systems. Future research will focus on integrating generative models like GANs for realistic data synthesis and developing comprehensive open-source benchmarks for Indian script recognition.

Keywords: *Hindi Script Recognition, Scene Text Recognition, Handwriting Recognition, Synthetic Data, Transfer Learning, Multilingual Models*.

1. Introduction

The ability to accurately recognize text from images, commonly known as **Scene Text Recognition (STR)**, and handwritten documents, referred to as **Handwriting Recognition (HWR)**, is an essential aspect of advancing **Optical Character Recognition (OCR)** technologies. For globally dominant languages like English, significant progress has been achieved with the help of large datasets and advanced deep learning models. However, the same level of progress has not been realized for resource-scarce languages such as Hindi, which is widely spoken in India and written in the Devanagari script [1]. Hindi's rich script structure, featuring complex conjuncts, diverse ligatures, and variable character forms, poses unique challenges to automated text recognition systems. Additionally, the diversity in font styles, non-standardized handwriting, and varying scene backgrounds further complicate the recognition process. These challenges underscore the pressing need for tailored solutions to enhance STR and HWR performance specifically for Hindi. One of the critical limitations in developing robust recognition systems for Hindi is the **lack of large, annotated datasets** that adequately capture the variations in font types, handwriting styles, and environmental conditions in which the text appears. Unlike English and other [2] Latin-based scripts, where comprehensive datasets are publicly available, Hindi suffers from data scarcity, especially in scene text and handwritten formats. This scarcity hampers the ability to train deep learning models effectively, resulting in underperforming systems when applied to real-world scenarios. Moreover, conventional models trained on Latin scripts do not generalize well to Hindi due to the scriptural and linguistic differences, leading to sub-optimal recognition outcomes.

To overcome these limitations, researchers have increasingly turned to **synthetic data generation** techniques. Synthetic datasets, created using a variety of fonts, text distortions, and background conditions, simulate real-world variability and help in enhancing model robustness [3]. For Hindi, generating synthetic data that includes both Unicode and non-Unicode fonts, as well as diverse handwriting samples, is pivotal. These synthetic datasets serve as a vital supplement to the limited real-world data, enabling the training of deep neural networks capable of handling script diversity and environmental noise. Additionally, such datasets facilitate the inclusion of underrepresented character combinations and styles that might not be abundantly available in naturally occurring data [4].

Transfer learning emerges as another powerful strategy to address the challenges associated with Hindi text recognition. Transfer learning allows models pre-trained on large datasets in resource-rich languages to be fine-tuned on Hindi-specific data, effectively transferring learned representations across scripts. This method has shown promising results in various NLP [5] and computer vision tasks by reducing the dependence on large labeled

1Assistant Professor, Department of Computer Science & Engineering, Institute of Engineering & Science, IPS Academy, Indore, India.
Email Id: barkhasahu@ipsacademy.org
Corresponding Author: Barkha Sahu
Email Id: barkhasahu@ipsacademy.org

datasets. In the context of Hindi STR and HWR, transfer learning from models trained on multiple Indian scripts or multilingual corpora can enhance performance, as scripts like Bengali and Marathi share structural similarities with Devanagari. This cross-script learning approach leverages common features across languages, making the recognition models more adaptable and accurate for Hindi [6].

Recent studies have demonstrated that incorporating **non-Unicode fonts and diverse handwriting styles** in synthetic datasets, combined with transfer learning, significantly boosts the recognition accuracy for Hindi. [7] For instance, experiments on the IIIT-ILST Hindi dataset showed a remarkable **33% improvement in Word Recognition Rate (WRR)** when models were trained with such enriched data. Furthermore, transfer learning experiments across six Indian languages revealed that Hindi models benefited more from related scripts than from English-based datasets. This insight highlights the potential of developing **multilingual and multi-script recognition frameworks** tailored for Indian languages [8].

In summary, while significant advancements have been made in STR and HWR for Latin scripts, Hindi remains a challenging yet vital frontier in OCR research. Addressing these challenges requires a combined approach of synthetic data generation, transfer learning, and cross-lingual model training. The development of robust and accurate recognition systems for Hindi [9] is not just a technical necessity but also a cultural imperative, as it enables the digitization and accessibility of a vast corpus of printed and handwritten content in the Hindi language. Future work in this domain can further explore **Generative Adversarial Networks (GANs)** for more realistic data synthesis, integrate context-aware models like transformers, and establish open-source benchmarks to propel research in Hindi script recognition to global standards [10].

2. Literature review

The Hindi Named-Entity Recognition (NER) system has evolved through models like BL-MENE and CP-MENE. While the baseline BL-MENE suffers from boundary detection and misclassification issues, the CP-MENE framework addresses these by incorporating additional features like part-of-speech, gazetteers, and relative pronouns. CP-MENE leverages recursive relationships and regular expressions to refine pattern extraction, significantly improving NE recognition on the Hindi health dataset (HHD) sourced from Kaggle, across Person, Disease, Consumable, and Symptom categories (Jain et al., 2022) [1].

For Hindi handwriting recognition, a hybrid MLPNN/HMM model was introduced, combining Hidden Markov Models with multilayer perceptrons. Input signals are transformed into segments via an elliptical method for feature extraction. The model, trained on DBMs Hindi database, achieved a recognition accuracy of 97.8%, outperforming prior systems. This demonstrates the potential of combined neural networks and HMMs for improving character recognition accuracy in Hindi script, a significant advancement over traditional online recognition techniques (Patil & Aithal, 2022) [2].

Automatic text summarization (ATS) for Hindi, utilizing Real Coded Genetic Algorithm (RCGA), efficiently processes health corpus data through phases like preprocessing, feature extraction, and sentence ranking. The RCGA optimizes feature weights using genetic operators to enhance summary quality, evaluated via ROUGE metrics. The method outperformed existing summarizers, achieving up to 65% reduction while maintaining relevance and coherence, indicating RCGA's efficacy in handling feature-rich Hindi text summarization tasks (Jain et al., 2022) [3].

Word Sense Disambiguation (WSD) in Hindi is enhanced using a genetic algorithm that interprets ambiguous nouns based on surrounding context. By utilizing dynamic configuration windows, the approach considers adjacent terms for accurate disambiguation. This model achieves an 80% accuracy, surpassing other methods like graph-based and probabilistic models, by addressing Hindi's linguistic complexities and limited resources for precise WSD (Bhatia et al., 2022) [4].

The BERT model, renowned for contextual language understanding, was adapted for Hindi to address limitations in multilingual NLP tools. Trained on Hindi Wikipedia, this system enhances document retrieval based on similarity scores for Hindi queries. This effort bridges the gap in Hindi-centric NLP models, as most prior solutions predominantly focus on English or resource-rich languages, leaving Indian languages underrepresented (Rajeshwari & Kallimani, 2022) [5].

A unique approach for Hindi poetry classification was introduced by leveraging Rasa-based emotional classification. Using lexical features, Hindi WordNet, Latent Semantic Indexing, and Word2Vec embeddings, poems were categorized according to mood and themes. With a repository of 37,717 poems, the system showed reliable performance on a dataset of 945 poems, highlighting the effectiveness of SVM classifiers in cultural and literary text classification tasks (Prakash et al., 2022) [6].

An automated system for extracting multiword expressions (MWEs) from Hindi and Urdu corpora was developed using part-of-speech tagging, pattern matching, and the CRF++ model. This system achieved high extraction accuracy (96.82% for Hindi and 96.62% for Urdu), aiding in better linguistic resource building for machine translation and NLP applications where idiomatic expressions play a crucial role (Gupta & Joshi, 2022) [7].

Scene text recognition in Indian languages benefits from transfer learning and font diversity strategies. Incorporating non-Unicode fonts alongside Unicode fonts, the system enhanced recognition accuracy across languages like Hindi, Tamil, and Gujarati. Results showed significant improvements, particularly on the IIIT-ILST Hindi dataset, with a Word Recognition Rate increase of over 33%, demonstrating the potential of synthetic data augmentation for multilingual STR tasks (Gunna et al., 2022) [8].

Handwritten Hindi character recognition using SVM combined with morphological features, HOG, and edge detection achieved a 96.97% accuracy rate. The summarization component ranks sentences based on statistical features and SVM optimization, achieving precision of 72% at 50% compression and 60% at 25% compression. This dual approach of recognition and summarization advances Hindi text processing, previously underexplored compared to English counterparts (Dhankhar et al., 2022) [9].

Context-based translation for Out Of Vocabulary (OOV) words in Hindi-English CLIR enhances retrieval performance. Utilizing large unlabeled corpora and a small bilingual corpus, the proposed method improved Recall and MAP significantly over traditional SMT, reducing OOV incidence by up to 1.73% in FIRE 2011 dataset evaluations, thereby refining Hindi-English cross-lingual translations (Sharma et al., 2022) [10].

A Hindi language-specific question answering framework based on string matching was developed to enhance web-based knowledge retrieval for direct and concise queries. Tested on various query types, the system achieved 93.33% accuracy, illustrating the viability of such tailored approaches for better information accessibility in Hindi, a frequently underserved language online (Mehta et al., 2022) [11].

A Sanskrit to Hindi machine translation system employing hybrid direct and rule-based processing was introduced, integrating Elasticsearch and parse tree construction. Using bilingual dictionaries and grammatical corpora, the system achieved a BLEU score of 51.6%, demonstrating improved handling of linguistic divergences between Sanskrit and Hindi compared to existing models (Sethi et al., 2022) [12].

An innovative Hindi CAPTCHA system combining printed and handwritten characters was designed to enhance security and usability. After breaking traditional CAPTCHA schemes, the new model achieves 100% resistance against computer attacks and 90% user usability, addressing vulnerabilities in Hindi language CAPTCHAs (Kumar et al., 2022) [13].

Addressing string manipulation challenges in Indian languages, a framework for consistent text processing on the web was proposed, focusing on the orthographic variations in Hindi. This initiative builds on W3C models for string matching, aiming to standardize handling of diverse encodings for better Indian language representation online (Verma et al., 2021) [14].

Three computational tools—Text2Mātrā, RPaGen, and FoSCal—were developed for analyzing Hindi poetry, mapping poetic elements like metre, rhyme patterns, and figures of speech numerically. This is the first system to quantify literary aesthetics in Hindi, facilitating research in education, literary criticism, and philology (Naaz & Singh, 2022) [15].

A document vector embedding-based ATS system was proposed for Hindi and English, enhancing summarization by emphasizing redundancy, diversity, and compression. Outperforming baselines like TextRank and LSA, the model achieved macro-average F-scores of 18.5% (Hindi) and 26% (English), marking a significant step in multilingual summarization research (Rani & Lobiyal, 2022) [16].

An RNN-based POS tagger for Hindi was developed, integrating LSTM and comparing against methods like HMM, SVM, and decision trees. The approach demonstrated improved tagging accuracy, supporting better syntactic parsing for low-resource languages like Hindi, which is crucial for advanced NLP applications (Mishra et al., 2022) [17].

Shabd, a psycholinguistic database for Hindi, compiles word frequencies from a 1.4 billion word corpus, enhancing lexical decision tasks compared to smaller corpora. With comprehensive datasets of words and bigrams, Shabd aids linguistic research by offering reliable frequency measures and contextual diversity metrics (Verma et al., 2022) [18].

For Hindi poetry translation into English, a hybrid MT method combining rule-based and statistical techniques was proposed, addressing the semantic and syntactic nuances unique to Hindi poetry. The model outperformed Google and Microsoft translators, showcasing better disambiguation and context retention (Chakrawarti et al., 2022) [19].

A deep neural network for Hindi NER was proposed, combining CNN, Bi-LSTM, and CRF models with character and word embeddings. This hybrid architecture enhances recognition accuracy for resource-scarce Hindi, effectively handling out-of-vocabulary issues (Sharma et al., 2020) [20].

In the WAT2022 multimodal translation challenge, a transliteration-based phrase augmentation improved English-Hindi translation. The team's approach secured second place with a BLEU score of 39.30, validating the benefit of integrating text and visual data in low-resource translation tasks (Laskar et al., 2022) [21].

A statistical sentence scoring method was introduced for Hindi extractive summarization, evaluating sentences across nine features. Applied to 2000 FIRE documents, the model's summaries were assessed using ROUGE at 40% retention, demonstrating efficiency in Hindi text summarization (Dhankhar & Gupta, 2022) [22].

A bilingual image document classification system was developed, distinguishing English and Romanized Hindi using SVM and random forest. Leveraging pseudo-thesaurus-based keyword identification, the system effectively classified and extracted information from bilingual documents, enhancing document processing in multi-script scenarios (Puri, 2022) [23].

An empirical study on phrase-based SMT for English-Hindi translation was conducted, analyzing the impact of reordering and language models. BLEU and TER metrics confirmed the system's effectiveness in navigating grammatical complexities between the two languages, essential for improving translation quality (Babhulgaonkar & Sonavane, 2022) [24].

3. Research Gaps Identified

1. **Resource Limitation for Hindi NLP Tasks :** While significant strides have been made in NLP for resource-rich languages, Hindi still suffers from a lack of comprehensive datasets, annotated corpora, and linguistic resources. This limitation is evident in tasks like Named Entity Recognition (NER), Part-of-Speech (POS) tagging, and Word Sense Disambiguation (WSD), where existing models underperform due to insufficient training data and domain-specific corpora [1][4][17].

2. **Inadequate Multimodal and Multilingual Models :** Most NLP advancements, including BERT and multilingual models, predominantly focus on English or major world languages. There is a scarcity of efficient multimodal translation systems specifically designed for Hindi or cross-lingual applications involving Hindi. The underrepresentation of Hindi in multimodal datasets hinders progress in robust translation and retrieval tasks [5][21].

3. **Limited Research in Hindi Literary Analysis and Poetic Structures :** Although some tools for poetry analysis in Hindi exist, they are limited to structural elements like rhyme patterns and metrical analysis. There's a notable gap in developing advanced computational models capable of capturing deep semantic, emotional, and cultural nuances in Hindi literary texts, especially poetry [6][15][19].

4. **Underdeveloped Scene Text and Handwriting Recognition :** Current scene text recognition (STR) and handwriting recognition methods for Hindi lag behind those for Latin-based scripts due to font diversity, script complexity, and limited annotated datasets. Transfer learning strategies have shown potential, but robust, generalized models for varied Hindi scripts remain underdeveloped [8][9].

5. **Deficiencies in Automated Summarization for Hindi :** Most Automatic Text Summarization (ATS) models show promise but are often adapted from English-centric algorithms without sufficient optimization for Hindi's linguistic structures. There is also limited exploration of abstractive summarization for Hindi, with extractive methods still dominating the field [3][16][22].

6. **Gaps in Robust Information Retrieval and Question Answering Systems :** Although efforts have been made in developing Hindi question answering and document retrieval systems, they often rely on basic string matching and lack semantic understanding. Advanced retrieval models, leveraging deep contextual embeddings specifically trained on Hindi, are still an open research area [5][11][14].

7. **Challenges in Machine Translation for Complex Hindi Structures :** Machine translation systems, especially Statistical Machine Translation (SMT) and Neural Machine Translation (NMT), struggle with Hindi due to grammatical complexity and divergence from languages like English. Additionally, the lack of domain-specific bilingual corpora limits translation quality for niche content such as poetry or specialized domains [12][19][24].

8. **Security and Usability in Hindi-based CAPTCHAs :** CAPTCHA systems in Hindi are still in nascent stages, with limited research on creating designs that balance both security and user accessibility. The fusion of printed and handwritten text has been proposed, but further research is needed to improve adaptability and robustness against evolving security threats [13].

9. **Underexplored Potential of Cross-Lingual Information Retrieval (CLIR) :** Although context-based translation methods for OOV words in CLIR have improved, Hindi still lacks sophisticated CLIR systems that seamlessly handle OOV challenges and ensure high retrieval accuracy in cross-lingual settings [10].

10. **Inconsistent Standardization for String Processing in Indian Languages :** There is a lack of standardized methodologies for string processing, matching, and searching in Indian languages like Hindi on the web, which creates barriers for consistent multilingual data handling and digital literacy tools [14].

4. Solutions to Address the Identified Research Gaps

1. **Develop Comprehensive and Annotated Datasets for Hindi :** To overcome data scarcity, dedicated efforts should focus on building large, diverse, and domain-specific annotated datasets for Hindi NLP tasks. Initiatives like community-driven corpus building, government-backed data repositories, and crowd-sourcing annotations can enhance the quality and quantity of datasets for NER, POS tagging, WSD, and sentiment analysis. Leveraging transfer learning from multilingual models can further strengthen model training on limited Hindi data.

2. **Advance Multimodal and Cross-Lingual Models :** Design specialized multimodal translation systems integrating text, speech, and visual inputs for Hindi and other Indian languages. Cross-lingual pretraining frameworks can be extended to develop Hindi-centric models, enhancing translation accuracy and retrieval in multilingual contexts. Incorporating transliteration and domain adaptation strategies can improve translation and cross-lingual information retrieval performance.

3. **Develop Deep Semantic Models for Hindi Literature and Poetry Analysis :** Innovative deep learning architectures, such as transformer-based models, should be tailored for analyzing Hindi literary texts. Models that capture cultural, emotional, and semantic depths can enable automated thematic and sentiment

classification of poetry and prose. Embedding cultural context into training data can enrich interpretations of complex literary forms.

4. **Enhance Scene Text and Handwriting Recognition for Hindi :** Building extensive annotated datasets with diverse fonts, scripts, and handwriting styles will support the development of robust STR and handwriting recognition models. Adopting advanced techniques like Generative Adversarial Networks (GANs) for synthetic data augmentation and transfer learning can improve recognition accuracy across varied Hindi scripts and real-world conditions.
5. **Innovate Abstractive Summarization Techniques for Hindi :** Shift focus from extractive to abstractive summarization by developing models based on advanced encoder-decoder architectures like T5 or mBART, customized for Hindi. Fine-tuning these models on Hindi corpora can facilitate the generation of coherent, context-rich, and semantically accurate summaries that better reflect the nuances of the language.
6. **Design Advanced Semantic Retrieval and QA Systems :** Develop deep learning-based question answering and information retrieval systems that utilize Hindi-trained embeddings (like IndicBERT or MuRIL). Semantic search techniques and contextual embeddings can replace traditional keyword-based retrieval, enhancing accuracy in understanding and answering queries in Hindi.
7. **Improve Machine Translation Quality for Complex Hindi Structures :** Adopt hybrid translation approaches that combine rule-based, statistical, and neural methods to better handle the syntactic and morphological complexities of Hindi. Creating domain-specific bilingual corpora, especially for specialized content like legal documents or poetry, will further improve translation quality and contextual accuracy.
8. **Strengthen Hindi CAPTCHA Security and Usability :** Introduce dynamic CAPTCHA generation methods using AI-driven variations in fonts, colors, and styles for Hindi. Incorporating adversarial testing frameworks can help identify and patch vulnerabilities, ensuring both robust security against bots and high usability for human users.
9. **Advance Cross-Lingual Information Retrieval (CLIR) Techniques :** Enhance CLIR systems by integrating contextual embeddings and leveraging large multilingual pre-trained models for better handling of OOV words. Developing adaptive algorithms that dynamically learn from new corpora can improve precision and recall in cross-lingual searches involving Hindi.
10. **Establish Standardized Protocols for String Processing in Hindi :** Collaborate with standardization bodies like W3C to formalize string processing rules tailored for Indian languages. Creating open-source libraries that support efficient string matching, indexing, and searching across varied encodings and orthographic variations will ensure consistency and reliability in web-based and computational applications for Hindi.

5. Conclusion & Future Work

5.1 Conclusion : This research emphasizes the significant advancements achievable in **Scene Text and Handwriting Recognition (STR and HWR)** for Hindi through the integration of **synthetic data generation** and **transfer learning techniques**. By augmenting datasets with diverse font styles, scripts, and handwritten samples—including non-Unicode fonts—recognition models can better generalize across real-world scenarios. Our exploration confirms that models trained with enriched synthetic datasets exhibit improved accuracy and robustness, especially when combined with transfer learning from multilingual or script-similar datasets. These improvements address the historical underperformance of Hindi STR and HWR compared to Latin-based scripts. The findings also highlight that Hindi, with its rich script complexity, benefits significantly from tailored data augmentation and advanced deep learning models, setting a new benchmark for multilingual and multi-script recognition systems.

5.2 Future Work : Future research can explore **Generative Adversarial Networks (GANs)** to create even more realistic synthetic data, capturing intricate handwriting variations and rare script forms in Hindi. Additionally, developing **multilingual STR frameworks** that allow cross-lingual learning across Indian scripts (like Devanagari, Bengali, Tamil) can further enhance recognition accuracy. Integrating **contextual understanding models**, such as transformers, could improve not only character-level accuracy but also word and sentence-level comprehension in noisy or complex visual backgrounds. Finally, building an open-source, large-scale Hindi STR and HWR dataset and benchmarking it with standardized protocols would greatly contribute to the NLP and computer vision communities, fostering further innovation in this underrepresented research area.

References

[1] Jain, Arti, Divakar Yadav, Anuja Arora, and Devendra K. Tayal. "Named-Entity Recognition for Hindi language using context pattern-based maximum entropy." *Computer Science* 23 (2022): 81-115.

[2] Patil, Vinita, and P. S. Aithal. "A Mixture of MLPNN/HMM to Demonstrate the Procedure for Online Hindi Writing Recognition." *International Journal of Case Studies in Business, IT and Education (IJCSBE)* 6, no. 1 (2022): 414-425.

[3] Jain, Arti, Anuja Arora, Jorge Morato, Divakar Yadav, and Kumar Vimal Kumar. "Automatic text summarization for Hindi using real coded genetic algorithm." Applied Sciences 12, no. 13 (2022): 6584.

[4] Bhatia, Surbhi, Ankit Kumar, and Mohammed Mutillah Khan. "Role of genetic algorithm in optimization of Hindi word sense

disambiguation." *IEEE Access* 10 (2022): 75693-75707.

[5] Rajeshwari, S. B., and Jagadish S. Kallimani. "Development of Optimized Linguistic Technique Using Similarity Score on BERT Model in Summarizing Hindi Text Documents." In *Innovative Data Communication Technologies and Application: Proceedings of ICIDCA 2021*, pp. 767-781. Singapore: Springer Nature Singapore, 2022.

[6] Prakash, Amit, Niraj Kumar Singh, and Sujan Kumar Saha. "Automatic extraction of similar poetry for study of literary texts: An experiment on Hindi poetry." *ETRI journal* 44, no. 3 (2022): 413-425.

[7] Gupta, Vaishali, and Nisheeth Joshi. "Identification and extraction of multiword expressions from Hindi & Urdu language in natural language processing." International Journal of Advanced Technology and Engineering Exploration 9, no. 91 (2022): 807.

[8] Gunna, Sanjana, Rohit Saluja, and Cheerakkuzhi Veluthemana Jawahar. "Improving scene text recognition for Indian languages with transfer learning and font diversity." *Journal of Imaging* 8, no. 4 (2022): 86.

[9] Dhankhar, Sunil, Mukesh Kumar Gupta, Fida Hussain Memon, Surbhi Bhatia, Pankaj Dadheech, and Arwa Mashat. "Support Vector Machine Based Handwritten Hindi Character Recognition and Summarization." *Computer Systems Science & Engineering* 43, no. 1 (2022).

[10] Sharma, Vijay Kumar, Namita Mittal, and Ankit Vidyarthi. "Context-based translation for the out of vocabulary words applied to hindi-english cross-lingual information retrieval." *IETE Technical Review* 39, no. 2 (2022): 276-285.

[11] Mehta, Shikha, Sakshi Gupta, Raashi Agarwal, Shrashti Trivedi, and Prajjwal Dubey. "String Matching Based Framework for Online Hindi Question Answering System." In *Iberoamerican Knowledge Graphs and Semantic Web Conference*, pp. 312-321. Cham: Springer International Publishing, 2022.

[12] Sethi, Nandini, Amita Dev, Poonam Bansal, Deepak Kumar Sharma, and Deepak Gupta. "Hybridization based machine translations for low-resource language with language divergence." *ACM Transactions on Asian and Low-Resource Language Information Processing* (2022).

[13] Kumar, Mohinder, Manish Kumar Jindal, and Munish Kumar. "Design of innovative CAPTCHA for hindi language." *Neural Computing and Applications* 34, no. 6 (2022): 4957-4992.

[14] Verma, Prashant, Vijay Kumar, and Bharat Gupta. "Indian Languages Requirements for String Search/comparison on Web." In *International Conference on Artificial Intelligence and Speech Technology*, pp. 210-214. Cham: Springer International Publishing, 2021.

[15] Naaz, Komal, and Niraj Kumar Singh. "Design and development of computational tools for analyzing elements of Hindi poetry." *IEEE Access* 10 (2022): 97733-97747.

[16] Rani, Ruby, and D. K. Lobiyal. "Document vector embedding based extractive text summarization system for Hindi and English text." *Applied Intelligence* 52, no. 8 (2022): 9353-9372.

[17] Mishra, Atul, Soharab Hossain Shaikh, and Ratna Sanyal. "Context based NLP framework of textual tagging for low resource language." *Multimedia Tools and Applications* 81, no. 25 (2022): 35655-35670.

[18] Verma, Ark, Vivek Sikarwar, Himanshu Yadav, Ranjith Jaganathan, and Pawan Kumar. "Shabd: A psycholinguistic database for Hindi." *Behavior Research Methods* 54, no. 2 (2022): 830-844.

[19] Chakrawarti, Rajesh Kumar, Jayshri Bansal, and Pratosh Bansal. "Machine translation model for effective translation of Hindi poetries into English." *Journal of Experimental & Theoretical Artificial Intelligence* 34, no. 1 (2022): 95-109.

[20] Sharma, Richa, Sudha Morwal, Basant Agarwal, Ramesh Chandra, and Mohammad S. Khan. "A deep neural network-based model for named entity recognition for Hindi language." *Neural Computing and Applications* 32, no. 20 (2020): 16191-16203.

[21] Laskar, Sahinur Rahman, Rahul Singh, Md Faizal Karim, Riyanka Manna, Partha Pakray, and Sivaji Bandyopadhyay. "Investigation of english to hindi multimodal neural machine translation using transliteration-based phrase pairs augmentation." In *Proceedings of the 9th Workshop on Asian Translation*, pp. 117-122. 2022.

[22] Dhankhar, Sunil, and Mukesh Kumar Gupta. "A statistically based sentence scoring method using mathematical combination for extractive Hindi text summarization." *Journal of Interdisciplinary Mathematics* 25, no. 3 (2022): 773-790.

[23] Puri, Shalini. "Image classification with information extraction by evaluating the text patterns in bilingual documents." In *International Conference on Advanced Network Technologies and Intelligent Computing*, pp. 115-137. Cham: Springer Nature Switzerland, 2022.

[24] Babhulgaonkar, Arun, and Shefali Sonavane. "Empirical analysis of phrase-based statistical machine translation system for English to Hindi language." *Vietnam Journal of Computer Science* 9, no. 02 (2022): 135-162.