IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 15, No. 2, April 2026, pp. 1771∼1782
ISSN: 2252-8938, DOI: 10.11591/ijai.v15.i2.pp1771-1782

Transformer-based Hindi image description and storytelling using enhanced attention and FastText embeddings

Anjali Sharma[1], Mayank Aggarwal[1], Jitin Khanna[2]
[1]Department of Computer Science and Engineering, Faculty of Engineering and Technology, Gurukula Kangri (Deemed to be University), Haridwar, India
[2]Manager Data and Analytics, IBM, Paramus, United States

Article Info

Article history:

Received Jun 23, 2025
Revised Feb 6, 2026
Accepted Mar 5, 2026

Keywords:

Evaluation metrics
FastText embeddings
Hindi image
Squeeze-and-excitation
Transformer models

ABSTRACT

This work presents a novel image description generation framework that combines a Transformer-based encoder-decoder architecture with a custom squeeze-and-excitation (SE) attention block integrated into an EfficientNet feature extractor. The decoder uses FastText embeddings specifically trained for Hindi and is evaluated on the Microsoft common objects in context (MS-COCO) dataset. To improve the captioning process, the model incorporates a generative pre-trained transformer (GPT) module to generate narrative descriptions based on the initial captions and applies multiple similarity metrics to assess output quality. The proposed system significantly outperforms existing methods, achieving high bilingual evaluation understudy (BLEU) scores (BLEU-1 to BLEU-4: 83.24, 73.17, 64.56, and 58.22), a consensus-based image description evaluation (CIDEr) score of 81.41, an F1 score of 90.29, and a metric for evaluation of translation with explicit ordering (METEOR) score of 81.18, indicating strong caption accuracy. Furthermore, the model achieves low error rates, with a word error rate (WER) of 15% and a character error rate (CER) of 11%. This work highlights the challenges of applying large-scale datasets like MS-COCO to resource-limited languages and demonstrates the effectiveness of integrating FastText embeddings with transformer-based models for Hindi image captioning.

Corresponding Author:

Anjali Sharma
Department of Computer Science and Engineering, Faculty of Engineering and Technology
Gurukula Kangri (Deemed to be University)
Haridwar, India
Email: 23631001@gkv.ac.in

1. INTRODUCTION

Image description synthesis employs perception techniques alongside language models to create correct and relevant text. Deep learning models like convolutional neural networks (CNNs), recurrent neural networks (RNNs), and architectures that are based on transformers (e.g., vision transformers (ViTs) and data-efficient image transformers (DeiTs)) have significantly advanced this field by producing semantically rich captions [1], [2]. Multilingual captioning, especially in complex languages like Hindi, faces challenges due to linguistic diversity and limited datasets. Hindi's unique syntax and morphology demand adapted models, but current resources like the translated Flickr8k dataset remain insufficient [3]–[5].

Despite progress in Hindi image captioning, limitations in dataset diversity and model adaptability hinder performance. Existing datasets like Flickr8k-Hindi and HIC restrict model generalizability. Integrating CNNs to extract visual features with transformer architectures for capturing global context enhances the overall quality of generated picture descriptions [6], [7]. Enhancing attention mechanisms further strengthens contextual and linguistic coherence [8]. This study targets three core objectives: building diverse datasets, refining contextual feature extraction, and improving linguistic accuracy.

This research advances Hindi image captioning by translating Microsoft common objects in context (MS-COCO) into Hindi, creating a robust dataset. It integrates squeeze-and-excitation (SE) attention-enhanced EfficientNet for detailed visual feature extraction and a transformer architecture tailored to Hindi's linguistic structure. FastText embeddings improve semantic richness, while generative pre-trained transformer (GPT) refines captions for narrative depth. Evaluated using bilingual evaluation understudy (BLEU), consensus-based image description evaluation (CIDEr), metric for evaluation of translation with explicit ordering (METEOR), word error rate (WER), and character error rate (CER), the model establishes a strong performance baseline. Unique contributions include a Devanagari-adapted SE block and GPT-based caption extension, addressing dataset scarcity and linguistic complexity, with broad implications for inclusive AI in low-resource, morphologically rich languages.

The proposed method surpasses prior Hindi image captioning approaches by combining SE-attention, EfficientNet, and FastText embeddings for detailed visual capture and narrative-level generation. Unlike earlier sentence-level models with limited multimodal fusion and evaluation, our approach introduces a linguistically rich framework and a custom Hindi MS-COCO dataset with comprehensive metric coverage. Table 1 shows the comparative analysis of Hindi image captioning approaches.

Table 1. Comparative analysis of Hindi image captioning approaches (✓ = present, x = absent)

Paper	1	2	3	4	5	6	7
Sharma *et al.* [9]	x	✓	✓	✓	x	x	✓
Kaur *et al.* [10]	x	✓	✓	✓	x	x	✓
Patel *et al.* [11]	x	✓	x	✓	x	x	x
Gupta *et al.* [12]	x	✓	✓	x	x	x	x
Mishra *et al.* [13]	x	✓	✓	✓	✓	x	✓
Mishra *et al.* [14]	x	✓	x	✓	✓	x	✓
Bisht *et al.* [15]	x	✓	✓	✓	✓	x	✓
Rai *et al.* [16]	✓	x	x	✓	✓	x	✓
Harshit *et al.* [17]	x	x	x	✓	✓	x	✓
Proposed method	✓	✓	✓	✓	✓	✓	✓

Note: 1: SE-attention, 2: CNN backbone, 3: embedding model, 4: multimodal integration, 5: dataset type MS-COCO, 6: narrative generation, and 7: evaluation metrics

Recent research in Hindi image captioning has explored various deep-learning architectures. Early efforts used CNN-long short-term memory (LSTM) models with datasets like Flickr8k and Flickr30k to generate Hindi captions, showing moderate success in visual-text alignment [18], [19]. Later works introduced attention blocks and transformer-based decoders (e.g., GPT-2), improving syntactic coherence and context capture [20], [13]. However, these models remain constrained by limited data and sentence-level generation, leaving gaps in narrative fluency and linguistic richness [21].

Attention mechanisms have become pivotal in enhancing caption quality by allowing models to focus on key image regions. Techniques like self-enhanced attention (SEA), top-down attention, and enhanced focal modules have demonstrated improved performance on standard datasets by refining spatial focus and object relevance [22]–[24]. Multiview and heterogeneous attention frameworks further advanced multimodal alignment and multilingual adaptability [25], [26], but often lacked customization for Indian language scripts like Devanagari. EfficientNet has emerged as a compelling image encoder, balancing accuracy and computational efficiency. Studies show its integration with transformer decoders improves feature extraction and caption fluency while maintaining low model complexity—an essential aspect for deployment in resource-constrained environments [27]–[29]. Lightweight combinations like EfficientNet-MobileNet-Transformer have proven effective across standard benchmarks [30], [31].

One of the most enduring issue in Hindi picture description generation is the language's intricate linguistic structure. Morphologically rich structures, high out-of-vocabulary (OOV) rates, and limited training

datasets hinder performance. FastText embeddings, with their subword modelling, have been shown to outperform traditional Word2Vec and even rival transformer-based embeddings for named entity recognition and sentiment tasks in Hindi [32], [33]. Additionally, lacking large-scale, diverse Hindi caption datasets like MS-COCO further restricts model generalizability. This research addresses these limitations through a novel integration of SE-attention EfficientNet, FastText embeddings, and transformer decoders—explicitly tailored to Hindi morphology and Devanagari script structure.

2. METHOD

This section presents a transformer-based Hindi image captioning framework combining EfficientNet-B4, SE-attention, and FastText embeddings to handle Hindi's morphological richness. The system is divided into four stages. These are data preprocessing, SE-attention feature extraction, transformer-based encoding-decoding, and GPT-based caption enhancement. Overall architecture: the model workflow (Figure 1) begins with translating MS-COCO captions into Hindi [34], cleaning and tokenizing them, and preparing image-caption pairs. Images are resized, converted into tensors, and passed to EfficientNet-B4, enhanced with SE blocks.

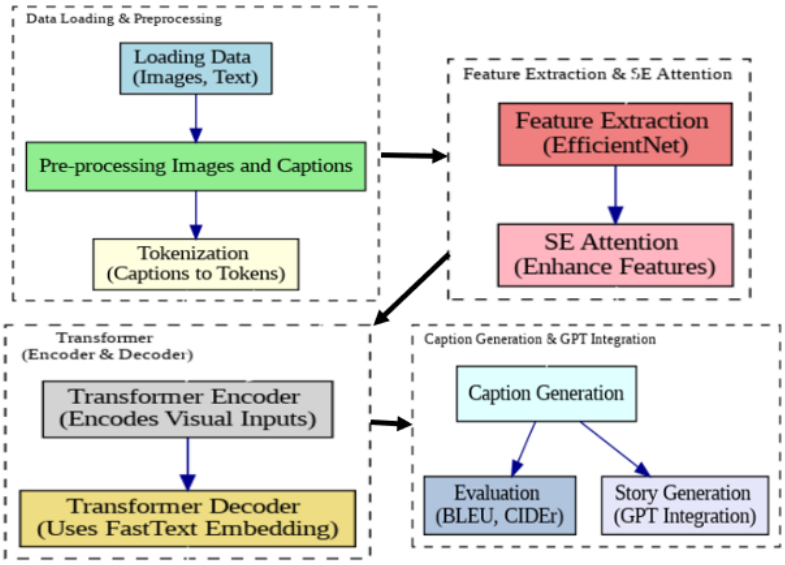


Figure 1. Proposed image captioning system

Feature extraction: EfficientNet + SE-attention. EfficientNet uses compound scaling for balancing depth, width, and resolution [35]. The SE-attention and encoder-decoder flowchart is shown in Figure 2, where Figure 2(a) shows the SE-attention mechanism and Figure 2(b) shows the encoder–decoder architecture. We extend its built-in SE block with a custom module, improving channel-wise and spatial recalibration.

$$X'i, j, c = s_c Xi, j, c \tag{1}$$

Transformer encoder-decoder with FastText: features are fed into a transformer encoder with FastText Hindi embeddings and positional encodings [36]. The decoder leverages self-attention, cross-attention, and sequential feed-forward networks to produce captions that are rich in contextual meaning.

$$\text{Attention}(Q_i, K_i, V_i) = \text{softmax}\left(\frac{Q_i K_i^T}{\sqrt{d_{head}}}\right) V_i \tag{2}$$

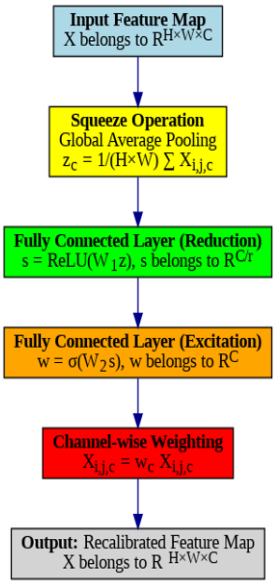
FastText's subword modeling effectively captures Hindi morphology and OOV words [37]. Its SGNS objective enhances rare word representation quality.

Caption generation and GPT integration: the decoder outputs Hindi captions.

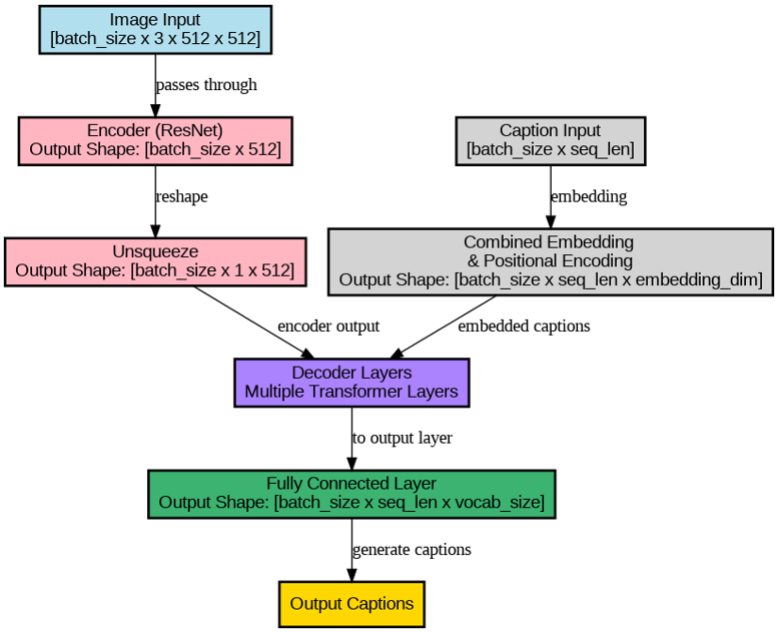
$$C_i' \leftarrow \text{Decode}(F_d) \tag{3}$$

Captions are evaluated using BLEU, CIDEr, METEOR [38]–[40], and error metrics WER, CER [41], [42]. We then refine them with GPT, improving fluency and narrative depth.

$$S_i \leftarrow \text{GPT}(C_i') \tag{4}$$



(a)



(b)

Figure 2. Attention and encoder–decoder flowcharts: (a) SE-attention mechanism and (b) encoder–decoder architecture flowchart

GPT was used as a post-processing module to enhance narrative richness while preserving semantic alignment. Transformer-generated captions were provided using a structured prompt with constraints on length, tense, and topic relevance, ensuring coherent and image-consistent storytelling. GPT-enhanced captions are

assessed using syntactic, lexical, semantic, and Jaccard similarity metrics [43], enabling broader applications like visual question answering and multilingual dialogue systems. In addition, a small-scale human evaluation assessed narrative coherence and expressive quality, confirming that GPT outputs improved fluency and descriptive depth without semantic drift. GPT was guided using a structured prompt template to control narrative generation and prevent hallucinated content. An example of the prompt used is:

> *Given the following Hindi image caption, generate a short, coherent narrative. Do not introduce new objects, actions, or events beyond the caption. Maintain the present tense and limit the output to 2–3 sentences.*

3. EXPERIMENTAL RESULTS

This section discuss the dataset, preprocessing pipeline, model training, performance metrics, ablation study, benchmarking, and qualitative evaluations.

3.1. Dataset details

We used the COCO-2017 dataset [34], translating English captions into Hindi using Google Translate API. Table 2 summarizes dataset stats; Figure 3 shows a sample translation frame. We translated COCO-2017 English captions into Hindi using Google Translate and applied a multi-stage quality assurance process to reduce noise and address Hindi's linguistic complexity.

Table 2. COCO-2017 dataset statistics

Dataset	Training	Validation	Testing	Vocabulary size
COCO-2017	118k	5k	40.7k	29,075

	image	caption	caption_hindi
0	/kaggle/input/coco-2017-dataset/coco2017/train...	A bearded man wearing a hat smiles and stands ...	टोपी पहने एक दाढ़ी वाला आदमी मुस्कुराता है और ...
1	/kaggle/input/coco-2017-dataset/coco2017/train...	A hand holding a Nintendo Wii game controller.	एक हाथ में निनटेंडो Wii गेम कंट्रोलर थामे हुए।
2	/kaggle/input/coco-2017-dataset/coco2017/train...	The small bathroom has a colorful shower curta...	छोटे बाथरूम में रंगीन शॉवर पर्दा है जिसमें पें...
3	/kaggle/input/coco-2017-dataset/coco2017/train...	A cat is cradled by a female while she is usin...	एक बिल्ली को एक मादा ने गोद में उठा रखा है जब ...
4	/kaggle/input/coco-2017-dataset/coco2017/train...	there is a plate of food on display on the table	मेज पर भोजन की एक प्लेट प्रदर्शित है

Figure 3. A sample translation frame

Translation quality assurance and validation: to improve translation quality, a three-stage post-translation validation process was applied.

- Automated filtering: captions were normalized using Unicode standardization for the Devanagari script, removal of duplicated tokens, punctuation correction, and elimination of non-Hindi artifacts.
- Semantic consistency: Hindi captions H_i were compared with their English counterparts E_i using FastText embeddings. Captions with cosine similarity below a threshold ($\tau = 0.65$) were discarded:

$$\text{Sim}(E_i, H_i) = \frac{E_i \cdot H_i}{\|E_i\|\|H_i\|} \tag{5}$$

- Manual validation: a random subset of 5,000 image–caption pairs was reviewed by native Hindi speakers, with over 93% of captions deemed linguistically acceptable after automated filtering.

3.2. Data pre-processing and model training configuration

Images were resized to 224×224 and normalized. Captions were cleaned, tokenized, and padded to 51 tokens. FastText Hindi embeddings (300-dim) were projected to 512-dim using:

$$E' = WE + b \tag{6}$$

Where E is the original 300-dimensional FastText embedding, W is a learnable weight matrix of shape (512×300), and b is a bias vector.

Model training was performed using the AdamW optimizer with label smoothing (0.1), KLDivLoss, a warmup scheduler with 4000 steps, and gradient clipping (norm =1.0). A teacher forcing strategy was adopted throughout the training. The hyperparameters and associated model performance metrics are provided in Table 3.

Table 3. Training hyperparameters and model performance metrics

Hyperparameter	Value	Performance metric	Value
Label smoothing	0.1	Model size	157.3 MB
Optimizer	AdamW	Trainable params	39.2 M
Learning rate	5×10^{-4}	Inference time	16.1 ms/image
Loss function	KLDivLoss	GPU usage	4450 MB
Warmup Steps	4000	FLOPs	0.84 GFLOPs
Gradient clipping	1.0		
Training strategy	Teacher forcing		

3.3. Performance and evaluation metrics

BLEU, CIDEr, METEOR, WER, and CER were used to evaluate captioning. GPT-refined captions were assessed using syntactic, lexical, semantic, and Jaccard similarities. Evaluation metrics such as BLEU and CIDEr primarily rely on n-gram overlap and may be sensitive to surface-level variations in morphologically rich languages like Hindi, where multiple valid inflected forms and flexible word order are common. As a result, these metrics can underestimate semantic correctness despite accurate visual grounding. METEOR, along with WER and CER, provides complementary insight by accounting for linguistic variation, word alignment, and error patterns specific to Hindi. Training metrics and system performance trends are shown in Figure 4.

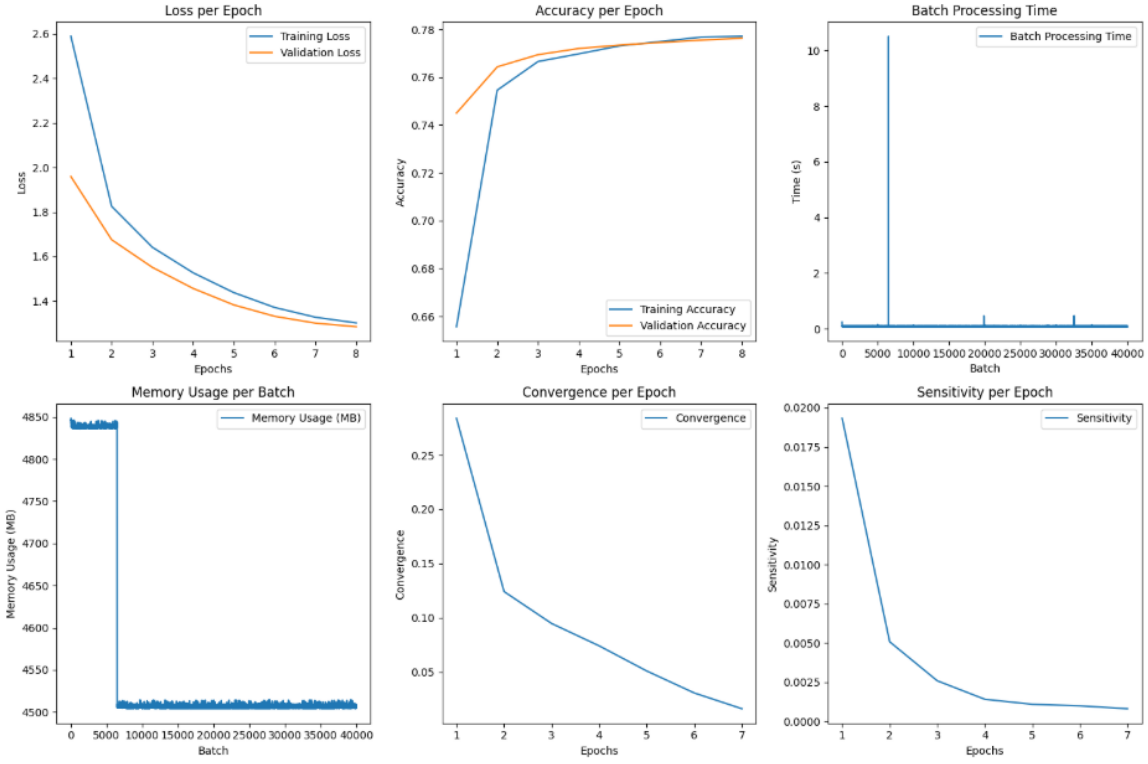


Figure 4. Hindi image captioning: training accuracy and loss overview

Final training and validation accuracy reached 76% and 78% respectively (Table 4). Model is efficient (16.1 ms inference time), using 4450 MB GPU memory. Evaluation metrics are presented in Figure 5, where Figure 5(a) shows image captioning evaluation metrics, Figure 5(b) shows WER and CER evaluation, Figure 5(c) shows scalability analysis, and Figure 5(d) shows ablation study. Table 4 summarizes the scores.

Table 4. Combined summary of model training, performance, and evaluation metrics

Metric	Training	Validation	Model metric	Value	Evaluation metric	Score
Final loss	1.3	1.2	Model size	157.3 MB	BLEU-1 to 4	83.24%, 73.17%, 64.56%, 58.22%
Final accuracy	76%	78%	Trainable params	39.2M	CIDEr	81.41%
–	–	–	Inference time	16.1 ms/image	METEOR	81.18%
–	–	–	GPU usage	4450 MB	F1-score	90.29%
–	–	–	FLOPs	0.84 GFLOPs	WER	14.82%
–	–	–	–	–	CER	10.75%

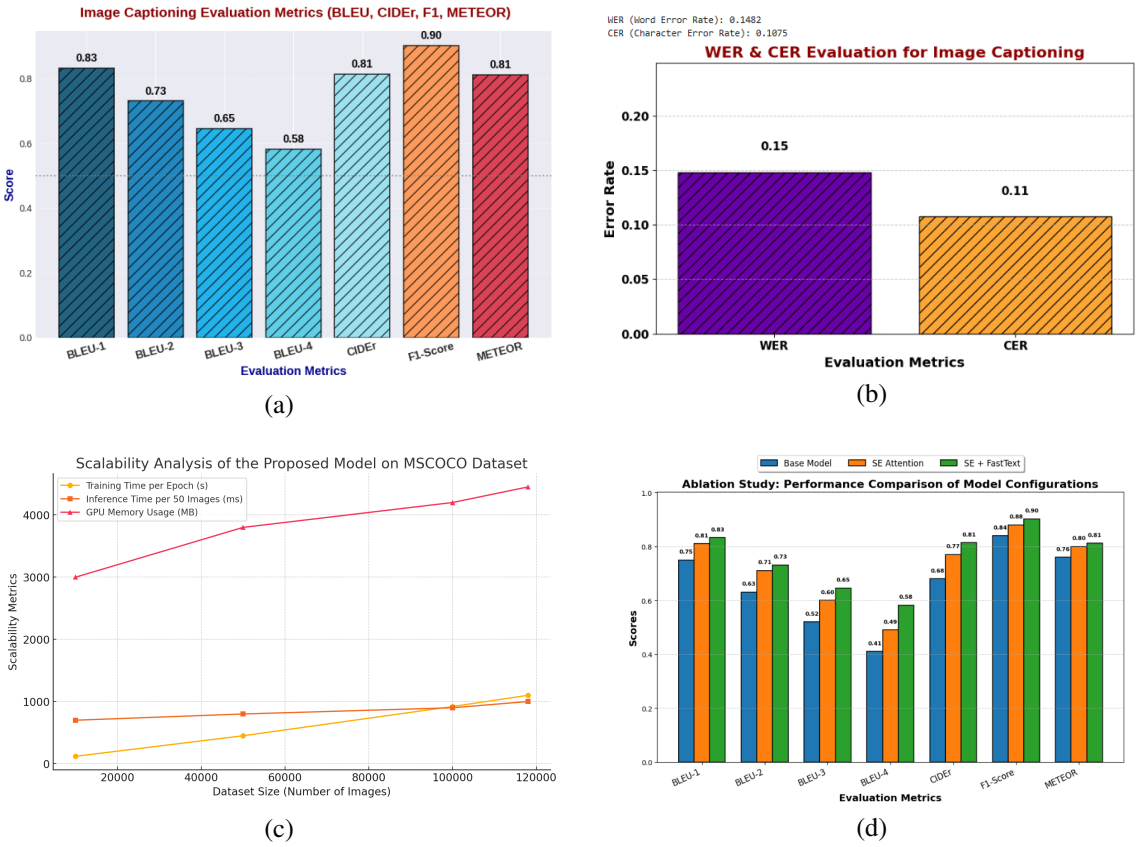


Figure 5. Combined view of four image captioning analysis visuals: (a) image captioning evaluation metrics, (b) WER and CER evaluation for image captioning, (c) scalability analysis, and (d) ablation study

3.4. Ablation analysis

Three model variants were tested: base transformer, with SE-attention, and with SE + FastText. Table 5 and Figure 5(d) demonstrate consistent improvement in all evaluation metrics with each architectural enhancement. To assess the statistical significance of performance gains introduced by SE-attention and FastText embeddings, a two-tailed paired t-test was conducted across five independent experimental runs (N=5). The test was applied to per-run evaluation scores computed on the same test set for all model configurations. Results indicate that the SE + FastText model achieves statistically significant improvements over the base configuration in BLEU-4, CIDEr, METEOR, and F1-score at the 95% confidence level ($p < 0.05$), suggesting that the observed gains are unlikely to be due to random variation.

Furthermore, as shown in Figure 5(c), the model is linearly scalable for training time and memory requirements as the dataset size increases from 10K to 118K samples. This confirms the model's suitability for large-scale deployment scenarios. The ablation study evaluates components that directly influence caption generation, including SE-attention and FastText embeddings. GPT was excluded from the ablation analysis, as it functions solely as a post-processing module for narrative enhancement and does not affect core captioning

metrics. Its impact is assessed separately through narrative-level evaluation. All ablation experiments were conducted over multiple runs, and the reported results reflect consistent performance trends across configurations, indicating the robustness of the observed improvements.

Table 5. Ablation study results with statistical significance analysis ($p < 0.05$)

Metric	Base	SE	SE+FastText	t-value	p-value
BLEU-4	41.08	49.44	58.22	3.12	0.021
CIDEr	68.49	77.26	81.41	3.45	0.015
F1	84.00	88.47	90.29	2.87	0.028
METEOR	76.23	80.76	81.18	2.41	0.041

3.5. Scalability and generalization

Beyond scalability evaluation, the proposed model was also tested on an external Hindi image captioning dataset, Flickr8k-Hindi, to examine its generalization capability. Using the same model architecture and evaluation setup, the approach maintained stable performance across datasets, obtaining a BLEU-4 score of 54.10, a CIDEr score of 74.85, a METEOR score of 76.32, and an F1-score of 87.45. Linguistic accuracy remained strong, with a WER of 18.6% and a CER of 13.9%. Although the dataset differs in scale and annotation style, the model consistently preserved relative performance gains over baseline configurations, demonstrating effective robustness and cross-dataset generalization.

3.6. Benchmarking and qualitative evaluation

The Table 6 proposed model outperforms existing Hindi image captioning approaches across BLEU scores, establishing a strong performance baseline. To assess the impact of GPT-based narrative enhancement, a comparative evaluation was conducted between base captions and GPT-enhanced narratives. Automatic similarity metrics, including syntactic, lexical, semantic, and Jaccard similarity, were used to verify semantic consistency and prevent content drift.

Table 6. Comparison of Hindi image captioning models

Authors	Model	B1	B2	B3	B4
Mishra *et al.*	Encoder-decoder	62.9	43.3	29.1	19.0
Singh *et al.*	CNN + RNN	51.3	30.4	16.7	12.4
Dhir	CNN + GRU	57.0	39.0	26.4	17.3
Rathi	CNN + LSTM	58.0	47.0	39.0	35.0
Meghwal	CNN + LSTM	62.5	45.8	32.8	23.2
Proposed model	CNN + transformer	83.24	73.17	64.56	58.22

In addition, a small-scale human evaluation was performed on a randomly selected subset of samples. Native Hindi speakers rated both versions using a three-point Likert scale (low, medium, high) based on narrative coherence and expressive richness. As summarized in Table 7, GPT-enhanced narratives consistently improved coherence and expressive depth while preserving the original semantic content. Sample qualitative results with captions and narrations are presented in Figure 6.

To contextualize the quantitative evaluation metrics, a qualitative error analysis was conducted on a randomly selected subset of 200 generated captions. Each caption was manually examined and assigned a dominant linguistic error category. The analysis revealed that most inaccuracies were minor and linguistically driven rather than semantic. Common error types included gender and number agreement mismatches, variations in word order, and postposition usage. These errors generally preserved the intended meaning but negatively affected n-gram–based metrics, highlighting the importance of complementing quantitative scores with qualitative analysis for morphologically rich languages such as Hindi. The results are summarized in Table 8.

Table 7. Human evaluation of caption vs. GPT-enhanced narrative

Criterion	Base caption	GPT-enhanced
Narrative coherence	Medium	High
Expressive richness	Low	Improved
Semantic consistency	High	High

Generated Caption: एक कुत्ते का सिर एक खिड़की के बाड़े से बाहर लटका हुआ है
Reference Caption: एक कुत्ते का सिर एक खिड़की से बाहर लटका हुआ है
BLEU-1: 0.8462, BLEU-2: 0.7966
BLEU-3: 0.7414, BLEU-4: 0.6703
CIDEr: 1.0000
WER: 0.1818, CER: 0.1957
F1-Score: 0.9091, METEOR: 0.8036

Sample Image - 1

Generated Caption:
ताजे हरे सेबों से भरा एक कटोरा रखा हुआ है|

Narration
सूरज की किरणों से चमकते हुए, एक सुंदर कटोरे में असली हरे सेबों की भरमार थी। उन सेबों का ताजगी से भरा हुआ स्वाद, मन को बहुत भाता था और आंखों को भी आनंद देता था। वे अपने आप में एक सुंदर दृश्य बनाते थे, जिसे देखकर किसी को भी मुस्कान आ जाती थी।

Syntactic (POS Tag) similarity: 0.87
Lexical similarity: 0.44
Semantic similarity: 0.60
Jaccard similarity: 0.45

Sample Image - 2

Generated Caption:
ब्रोकोली चावल मांस और अन्य सब्जियों की एक प्लेट|

Narration:
दिन का थकान दूर करने के लिए रसोई में खुशबू फैलाने वाली एक प्लेट ब्रोकोली, चावल, मांस और अन्य व्यंजनों की थाली तैयार की गई थी। उस प्लेट में रंग बिरंगी सब्जियाँ, गरमा गरम चावल और स्वादिष्ट मांस के टुकड़े अद्वितीय आकर्षण का स्रोत थे। वो थाली न केवल पेट को संतुष्ट करने के लिए थी, बल्कि आँखों को भी खुश करने वाली थी।

Syntactic (POS Tag) similarity: 0.82
Lexical similarity: 0.57
Semantic similarity: 0.62
Jaccard similarity: 0.51

Figure 6. Sample images with generated captions, evaluation scores, and GPT-enhanced narrations

Table 8. Distribution of common error types in Hindi caption generation

Error type	Percentage (%)
Gender/number agreement errors	36.0
Word order variations	28.5
Postposition usage errors	20.0
Lexical inflection errors	15.5

Scores indicate semantic correctness and fluency. GPT-based narrations further enhance richness and expressiveness and descriptions maintain semantic alignment with high syntactic and lexical similarity. The model effectively balances accuracy and creativity in Hindi caption generation.

4. CONCLUSION

This work presents an advanced Hindi image captioning framework that integrates a custom SE-attention mechanism with an EfficientNet-based transformer encoder–decoder architecture. Experimental results show substantial improvements in BLEU, CIDEr, and METEOR scores, along with reduced WER and CER, compared to existing methods. The use of FastText embeddings enables effective modeling of Hindi's morphological and syntactic characteristics, making the approach suitable for non-English and low-resource language captioning. The model maintains robust performance across datasets of varying scale. Although the model exhibits good cross-dataset generalization, domain-specific variations in visual content and linguistic style may affect performance in new application settings. Future work will focus on domain adaptation and transfer learning strategies, such as fine-tuning on target-domain data and multilingual pretraining, as well as exploring advanced attention mechanisms, multi-modal extensions, and reinforcement learning to further enhance caption quality.

FUNDING INFORMATION

This work was carried out independently and did not receive financial assistance from any governmental, corporate, or academic grant-awarding bodies.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Anjali Sharma	✓	✓	✓	✓	✓	✓		✓	✓	✓			✓	
Mayank Aggarwal		✓				✓		✓	✓	✓	✓	✓		
Jitin Khanna	✓		✓	✓		✓			✓		✓		✓	

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal Analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject Administration
Fu : **Fu**nding Acquisition

CONFLICT OF INTEREST STATEMENT

The authors declare that this research was conducted without any competing financial or personal interests.

ETHICAL CONSIDERATIONS

This study uses publicly available datasets and addresses ethical considerations related to automated image captioning and storytelling in Hindi. When applied to sensitive domains such as news or education, ensuring accuracy, cultural sensitivity, and human oversight is essential to prevent misinterpretation or misleading content.

INFORMED CONSENT

No informed consent was required, as the study did not involve human subjects or personal data.

DATA AVAILABILITY

The dataset employed in this research is publicly accessible and can be retrieved from the official COCO dataset website: https://cocodataset.org/#download.

REFERENCES

[1] K. Rage, "A study on different deep learning architectures on image captioning," in *2022 8th International Conference on Smart Structures and Systems (ICSSS)*, Apr. 2022, pp. 1–9, doi: 10.1109/ICSSS54381.2022.9782260.

[2] R. Castro, I. Pineda, W. Lim, and M. E. M. -Cayamcela, "Deep learning approaches based on transformer architectures for image captioning tasks," *IEEE Access*, vol. 10, pp. 33679–33694, 2022, doi: 10.1109/ACCESS.2022.3161428.

[3] J. Zhang, D. Guo, X. Yang, P. Song, and M. Wang, "Visual-linguistic-stylistic triple reward for cross-lingual image captioning," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 20, no. 4, pp. 1–23, Apr. 2024, doi: 10.1145/3634917.

[4] B. R. Reddy, S. Gunti, R. P. Kumar, and S. Sridevi, "Multilingual image captioning: multimodal framework for bridging visual and linguistic realms in Tamil and Telugu through transformers," *Research Square*, doi: 10.21203/rs.3.rs-3380598/v1.

[5] V. Jayaswal, R. Rani, and J. Kaur, "A deep learning-based efficient image captioning approach for Hindi language," in *Developments Towards Next Generation Intelligent Systems for Sustainable Development*. New York, United States: IGI Global, 2024, pp. 225–246, doi: 10.4018/979-8-3693-5643-2.ch009.

[6] H. Ahmadabadi, O. N. Manzari, and A. Ayatollahi, "Distilling knowledge from CNN-transformer models for enhanced human action recognition," in *2023 13th International Conference on Computer and Knowledge Engineering (ICCKE)*, Nov. 2023, pp. 180–184, doi: 10.1109/ICCKE60553.2023.10326272.

[7] W. Rine, R. Patel, and N. Steve, "Advanced language understanding with syntax-enhanced transformer," 2023, *Preprints*, doi: 10.20944/preprints202312.1673.v1.

[8] L. Shen *et al*., "Improving the robustness of transformer-based large language models with dynamic attention," in *Proceedings 2024 Network and Distributed System Security Symposium*, 2024, doi: 10.14722/ndss.2024.24115.

[9] S. Rawat, M. Manwal, and K. C. Purohit, "Simplifying Image captioning in hindi with deep learning," in *2024 International Conference on Computer, Electronics, Electrical Engineering & their Applications (IC2E3)*, Jun. 2024, pp. 1–7, doi: 10.1109/IC2E362166.2024.10827396.

[10] J. Kaur and G. S. Josan, "English to Hindi multi modal image caption translation," *Journal of scientific research*, vol. 64, no. 2, pp. 274–281, 2020, doi: 10.37398/JSR.2020.640238.

[11] S. De, R. Das, and A. S. Patel, "Bengali image caption generation using attention mechanism," in *2024 IEEE International Conference on Computer Vision and Machine Intelligence (CVMI)*, Oct. 2024, pp. 1–6, doi: 10.1109/CVMI61877.2024.10781602.

[12] R. Padate, A. Jain, M. Kalla, and A. Sharma, "Combining semi-supervised model and optimized LSTM for image caption generation based on pseudo labels," *Multimedia Tools and Applications*, vol. 83, no. 10, pp. 29997–30017, Sep. 2023, doi: 10.1007/s11042-023-16687-x.

[13] S. K. Mishra, S. Sinha, S. Saha, and P. Bhattacharyya, "Dynamic convolution-based encoder-decoder framework for image captioning in Hindi," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 22, no. 4, pp. 1–18, Apr. 2023, doi: 10.1145/3573891.

[14] S. K. Mishra, R. Dhir, S. Saha, P. Bhattacharyya, and A. K. Singh, "Image captioning in Hindi language using transformer networks," *Computers and Electrical Engineering*, vol. 92, Jun. 2021, doi: 10.1016/j.compeleceng.2021.107114.

[15] P. Bisht and A. Solanki, "Exploring practical deep learning approaches for English-to-Hindi image caption translation using transformers and object detectors," in *Applications of Artificial Intelligence and Machine Learning*, vol. 925, Singapore: Springer Nature, 2022, pp. 47–60, doi: 10.1007/978-981-19-4831-2_5.

[16] S. K. Mishra, G. Rai, S. Saha, and P. Bhattacharyya, "Efficient channel attention based encoder–decoder approach for image captioning in Hindi," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 21, no. 3, pp. 1–17, May 2022, doi: 10.1145/3483597.

[17] P. Bhardwaj, R. Tiwari, D. Thakur, and N. Jaglan, "Hate speech detection in Hinglish text using deep learning," B.Tech. project report, Jaypee University of Information Technology, Solan, India, 2023.

[18] A. K. Poddar and R. Rani, "Hybrid architecture using CNN and LSTM for image captioning in Hindi language," *Procedia Computer Science*, vol. 218, pp. 686–696, 2023, doi: 10.1016/j.procs.2023.01.049.

[19] A. Sethi, A. Jain, and C. Dhiman, "Image caption generator in Hindi using attention," in *Advances in Transdisciplinary Engineering*. Amsterdam, Netherlands: IOS Press, 2022, doi: 10.3233/ATDE220727.

[20] S. K. Mishra, S. Chakraborty, S. Saha, and P. Bhattacharyya, "GAGPT-2: a geometric attention-based GPT-2 framework for image captioning in Hindi," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 22, no. 10, pp. 1–16, Oct. 2023, doi: 10.1145/3622936.

[21] P. Singh, F. Raja, and H. Sharma, "Generating image captions in Hindi based on encoder-decoder based deep learning techniques," in *Reliability Engineering for Industrial Processes*, Cham, Switzerland: Springer, 2024, pp. 81–94, doi: 10.1007/978-3-031-55048-5_6.

[22] Q. Sun, J. Zhang, Z. Fang, and Y. Gao, "Self-enhanced attention for image captioning," *Neural Processing Letters*, vol. 56, no. 2, Apr. 2024, doi: 10.1007/s11063-024-11527-x.

[23] Y. Hua, P. Li, X. Zeng, and H. Xu, "Advanced techniques for Chinese image captioning: investigating attention mechanisms based on object detection for Chinese image caption generation," in *2024 5th International Conference on Machine Learning and Computer Application (ICMLCA)*, Oct. 2024, pp. 268–271, doi: 10.1109/ICMLCA63499.2024.10753950.

[24] A. Sattar, M. Assam, T. J. Alahmadi, U. A. Bhatti, H. Tang, and M. Aamir, "Remote sensing based advance image captioning improved feature attention," in *2024 7th International Conference on Pattern Recognition and Artificial Intelligence (PRAI)*, Aug. 2024, pp. 97–105, doi: 10.1109/PRAI62207.2024.10826910.

[25] Y. Fu, S. Fang, R. Wang, X. Yi, J. Yu, and R. Hua, "Multi-view attention with memory assistant for image captioning," in *2022 IEEE 6th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, Oct. 2022, pp. 436–440, doi: 10.1109/IAEAC54830.2022.9929571.

[26] Z. Song, Z. Hu, Y. Zhou, Y. Zhao, R. Hong, and M. Wang, "Embedded heterogeneous attention transformer for cross-lingual image captioning," *IEEE Transactions on Multimedia*, vol. 26, pp. 9008–9020, 2024, doi: 10.1109/TMM.2024.3384678.

[27] A. Joshi, A. Alkhayyat, H. Gunwant, A. Tripathi, and M. Sharma, "Enhancing image captioning performance based on efficientnet B0 model and transformer encoder-decoder," *AIP Conference Proceedings*, vol. 2919, no. 1, Mar. 2024, doi: 10.1063/5.0184395.

[28] U. A. A. Al-Faruq and D. H. Fudholi, "EfficientNet-Transformer for image captioning in Bahasa," *AIP Conference Proceedings*, vol. 2508, no. 1, Mar. 2023, doi: 10.1063/5.0118155.

[29] N. Wang *et al*., "Efficient image captioning for edge devices," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 2, pp. 2608–2616, Jun. 2023, doi: 10.1609/aaai.v37i2.25359.

[30] X. Zhang, M. Fan, and M. Hou, "Mobilenet V3-transformer, a lightweight model for image caption," *International Journal of Computers and Applications*, vol. 46, no. 6, pp. 418–426, Jun. 2024, doi: 10.1080/1206212X.2024.2328498.

[31] P. Bansal, K. Malik, S. Kumar, and C. Singh, "EfficientNet-based image captioning system," in *2023 International Conference on Device Intelligence, Computing and Communication Technologies, (DICCT)*, Mar. 2023, pp. 643–647, doi: 10.1109/DICCT56244.2023.10110117.

[32] R. Jha *et al*., "Image2tweet: Datasets in Hindi and English for generating tweets from images," in *Proceedings of the 18th International Conference on Natural Language Processing (ICON)*, Silchar, India, Dec. 2021, pp. 670–676.

[33] S. Rakshit, H. Dhawan, T. Gupta, and R. Narula, "An empirical comparison of Hindi-BERT and MuRIL for hate speech detection on social media platforms in Hindi language," in *Proceedings of the International Conference on Innovative Computing & Communication (ICICC 2024)*, 2024.

[34] T. -Y. Lin *et al*., "Microsoft COCO: common objects in context," in *Computer Vision – ECCV 2014*, vol. 8693, Cham, Switzerland, 2014, pp. 740–755, doi: 10.1007/978-3-319-10602-1_48.

[35] D. H. Fudholi, A. Zahra, and R. A. N. Nayoan, "A study on visual understanding image captioning using different word embeddings and CNN-based feature extractions," *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*, Feb. 2022, doi: 10.22219/kinetik.v7i1.1394.
[36] H. Chefer, S. Gur, and L. Wolf, "Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2021, pp. 397–406.
[37] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, Dec. 2017, doi: 10.1162/tacl_a_00051.
[38] K. Papineni, S. Roukos, T. Ward, and W. -J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in P*roceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, United States, Jul. 2002, pp. 311–318. doi: 10.3115/1073083.1073135.
[39] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: consensus-based image description evaluation," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, United States, 2015, pp. 4566-4575, doi: 10.1109/CVPR.2015.7299087.
[40] M. Puppala *et al.*, "METEOR: An enterprise health informatics environment to support evidence-based medicine," *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 12, pp. 2776–2786, Dec. 2015, doi: 10.1109/TBME.2015.2450181.
[41] M. Popović and H. Ney, "Word error rates: decomposition over POS classes and applications for error analysis," in *Proceedings of the Second Workshop on Statistical Machine Translation*, Prague, Czech Republic, Jun. 2007, pp. 48–55.
[42] D. K. Thennal, J. James, D. P. Gopinath, and M. A. K, "Advocating character error rate for multilingual ASR evaluation," in *Findings of the Association for Computational Linguistics: NAACL 2025*, Albuquerque, New Mexico, 2025, pp. 4926–4935, doi: 10.18653/v1/2025.findings-naacl.277.
[43] S. Bag, S. K. Kumar, and M. K. Tiwari, "An efficient recommendation generation using relevant Jaccard similarity," *Information Sciences*, vol. 483, pp. 53–64, May 2019, doi: 10.1016/j.ins.2019.01.023.

BIOGRAPHIES OF AUTHORS

Anjali Sharma is a dedicated academician with a strong background in Information Technology and Computer Science. She earned her B.Tech. in Information Technology and later completed her M.Tech. in Computer Science from the prestigious Dr. A. P. J. Abdul Kalam Technical University. She is pursuing her Ph.D. from Gurukul Kangri (Deemed University), further advancing her research expertise and academic pursuits. She can be contacted at email: 23631001@gkv.ac.in or awsanjali@gmail.com.

Mayank Aggarwal holds a Ph.D. in Computer Science. He is professor and dean, Faculty of Engineering and Technology, Gurukula Kangri (Deemed to be University), Haridwar. He has 20+ years of experience. He was a recipient of the gold medal and the University Topper in B.Tech., and published several research papers and imparted several trainings in the field of cloud computing in collaboration with IBM for students and faculties throughout the country. He can be contacted at email: mayank@gkv.ac.in.

Jitin Khanna is a seasoned techno-functional consultant with extensive experience in Business Intelligence and Analytics. He specializes in SAP S/4HANA embedded analytics, SAP BW/4HANA, SAP analytics cloud, and Microsoft power BI. With strong proficiency in designing scalable, cost-effective solutions, he collaborates closely with business analysts to translate functional requirements into technical specifications. He is a certified SAP sales and distribution consultant (SAP AG – Siemens), SAP HANA certified (openSAP), and ITIL foundation certified. His expertise spans SAP BODS, HANA modeling, SLT configuration, and industry solutions such as IS-utilities and IS-auto. He can be contacted at email: jitin.khanna1@ibm.com.