Research Paper

Real-Time Image Captioning for Visually Impaired Assistance Using Fine-Tuned Inception-ResNet Transfer Learning Model

Shanmugapriya Ganesan[1*], Palla Lalitha[2], Nirmala Devi Periyasamy[3], Mathamsetti Kaivalya[4], Julius Irudayasamy[5], Katanguri Swanthana[6]

1. *CSE Department, New Horizon College of Engineering, Bengaluru, Karnataka, India.*
2. *Department of CSE, Geetanjali College of Engineering and Technology, Hyderabad, Telangana, India.*
3. *Department of ECE, Kongu Engineering College, Perundurai, Erode, Vidya Nagar East, Tamil Nadu, India.*
4. *Department of ECE, Aditya University, Surampalem, Andhra Pradesh, India.*
5. *Department of English Language and Literature, College of Arts and Applied Sciences, Dhofar University, Salalah, Oman.*
6. *Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Bowrampet, Vaddeswaram, Andhra Pradesh, India.*

Citation Ganesan Sh, Lalitha P, Nirmala Devi P, Kaivalya M, Irudayasamy J, Swanthana K. Real-Time Image Captioning for Visually Impaired Assistance Using Fine-Tuned Inception-ResNet Transfer Learning Model. International Journal of Medical Toxicology and Forensic Medicine. 2026; 16:E52616.

https://doi.org/10.22037/ijmtfm.v16.52616

Article info:

Received: 27 June, 2026

Accepted: 09 July, 2026

Published: 16 September, 2026

Keywords:

Image captioning, Visually challenged people, Metaheuristic optimization, Feature extraction, Image pre-processing

ABSTRACT

Background: Visually challenged and blind people frequently face socio-economic barriers that can make it difficult for them to live independently and contribute fully to society. Currently, studies on assistive technology have advanced, leading to the development of a visual replacement for individuals with visual impairment. Image captioning is a task at the border between computer vision (CV) and natural language processing (NLP) that addresses making a textual explanation of the image. This paper presents a Metaheuristic Optimization-Based Image Captioning for Assisting Visually Challenged People Using Deep Learning (MOIC-AVCPDL) model to deliver image captioning tasks using advanced techniques.

Methods: The proposed MOIC-AVCPDL model employs InceptionResNetV2 for effective feature extraction, yielding rich, discriminative feature representations from the input data. The Bidirectional Gated Recurrent Unit technique is used for caption prediction. Furthermore, to maximize performance, the Multi-Objective Arithmetic Optimization Algorithm is utilized for hyperparameter tuning, enabling optimal parameter selection for improved accuracy.

Results: The proposed MOIC-AVCPDL achieved BLEU-1 scores of 78.08% on Flickr8K and 77.47% on Flickr30K, indicating improved image-captioning performance.

Conclusion: The proposed method delivers effective image captioning and has the potential to assist visually impaired people.

*** Corresponding Author:**
Shanmugapriya Ganesan, PhD
Department of Artificial Intelligence and Data Science, Adhiyamaan College of Engineering, Dr. M.G.R. Nagar, Hosur, Tamil Nadu, India.
E-mail: *spriyagsn@yahoo.com*

Introduction

Image captioning lies at the intersection of computer vision and natural language processing (NLP), generating a textual description of an image by using image processing, object identification, and understanding of the activities and scenes portrayed in it [1]. Image captioning has several applications in these fields: visual media retrieval, robotics, and accessibility. Image Captioning is a tool that assists blind persons in understanding and interpreting the contents of captured images [2]. The application empowers visually impaired users to capture and upload images of their surroundings and receive detailed descriptions in return. To overcome reliance on people, image captioning is essential, enabling independent functionality [3]. Assistive techniques are rehabilitative and adaptive systems that individuals use to execute tasks they have difficulty with or can't do without other people's help. Image captioning, united with the text-to-speech method, will support blind and visually impaired people [4].

Despite significant advancements in image captioning, producing precise and useful descriptions to assist visually impaired individuals remains challenging due to the complexity of scene understanding, semantic interpretation, and object relationships. Existing methods often suffer from limited caption quality, reduced contextual understanding, and suboptimal parameter selection that impact overall performance. To overcome these limitations, the study proposes a Metaheuristic Optimization-Based Image Captioning for Assisting Visually Challenged People Using Deep Learning (MOIC-AVCPDL) framework, an intelligent and reliable system developed to support visually challenged individuals. The main contribution of the proposed model is summarized below:

- Introduce InceptionResNetV2, a powerful hybrid convolutional neural network (CNN) model for feature extraction, which effectively identifies and extracts meaningful, hierarchical features from image data.

- Employ a bidirectional gated recurrent unit (BiGRU) for a reliable caption prediction process, enabling the learning of contextual dependencies by processing visual features in these two forward and backward directions.

- Integrate a multi-objective arithmetic optimization algorithm (MOAOA) for the hyperparameter tuning process, confirming the optimal selections to improve the accuracy of predictions.

Ayadi et al. [3] introduced a fresh hybrid paradigm of LSTM and CNN for Arabic handwritten character recognition (AHCR). Safiya and Pandian [4] presented an innovative framework for practical image captioning by the VGG16-LSTM DL model with CV support. Valipoor et al. [5] discussed the current challenges in CV-aided indoor activities in understanding assistive outcomes for visually disabled people. The authors [6] suggested a Red Deer Optimizer with an AI-based Image Captioning System (RDOAI-ICS) for visually challenged Individuals.

Ayadi et al. [7] proposed a novel system to improve the accuracy of handwritten Arabic recognition. The system is trained to rely on characteristics, namely style, size, and orientation, employing a conditional GAN structure. Gupta and Katal [8] proposed a hybrid CNN-RNN in which a multilayer CNN was used to extract an image's features, and an LSTM was used to generate understandable titles for images based on the language learned by the training network.

Materials and Methods

The proposed model follows a systematic approach to developing an image-captioning model to assist blind people. Figure 1 reveals the general framework of the MOIC-AVCPDL system. The proposed approach integrates advanced image preprocessing, deep feature extraction, caption prediction, and metaheuristic-driven hyperparameter optimization. The proposed model improves caption prediction performance while guaranteeing robustness and reliability. The proposed MOIC-AVCPDL model encompasses multiple stages, as shown in Figure 1.

Hybrid Feature Extraction Strategy

For feature extraction, the proposed MOIC-AVCPDL model uses InceptionResNetV2. The InceptionResNet method incorporates the inception structure and residual links [9]. The mathematical representation of the InceptionResNet-V2 is defined as demonstrated:

Stem Block

This system normally begins with the stem block for processing the input. The stem block's output Z_{stem} is provided as demonstrated:

$$Z_{stem} = StemBlock(x) \quad (1)$$

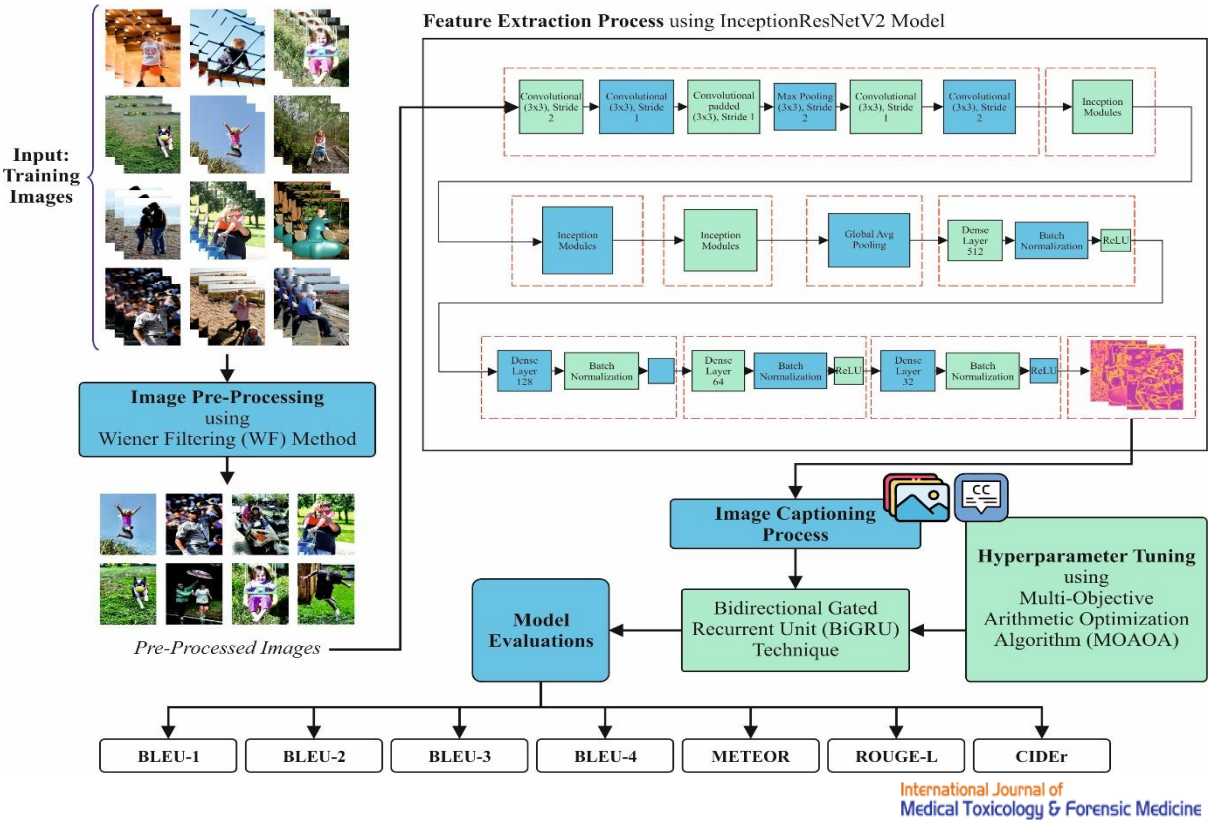

Figure 1. Overall architecture of the MOIC-AVCPDL approach.

Inception Modules

The main problem of InceptionResNet-V2 involves recurrent Inception units. All modules have numerous parallel divisions that process inputs in several ways, and their outputs must be related. In the given inception unit, the output $Z_{inception}$ can be given by,

$$Z_{inception} = InceptionModule(Z_{previous}) \quad (2)$$

Now, $Z_{previous}$ signifies the prior output block.

Residual Blocks

They are designed with skip connections that can be placed at different points in the system to simplify gradient flow during training. Like the presented residual block, the output $Z_{residual}$ was obtainable as demonstrated:

$$Z_{residual} = ResidualBlock(Z_{previous}) \quad (3)$$

Reduction Blocks

These blocks can be applied to reduce the spatial sizes of feature maps. Due to the provided block, the output $Z_{-reduction}$ was offered as demonstrated:

$$Z_{reduction} = ReduactionBlock(Z_{previous}) \quad (4)$$

Global Average Pooling (GAP)

It is also used to decrease the spatial sizes. The pooled output Z_{pooled} can be given as shown:

$$Z_{pooled} = GlobalAvgPool(Z_{previous}) \quad (5)$$

FC Layer

At last, an FC layer was used for classification. Such a layer is often paired with the Softmax activation function for multiclass classification. Assume that W_{fc} is the weight, and b_{fc} signifies the bias function; the last outcome has been demonstrated mathematically,

$$Y = Softmax(W_{fc} \cdot Z_{pooled} + b_{fc}) \quad (6)$$

Image Captioning using BiGRU Model

Following this, the BiGRU has been deployed for classification. The BiGRU Architecture is shown in Figure 2. The GRU has a better structure than the LSTM and is more efficient at capturing long-term dependencies and mitigating the vanishing gradient problem in RNNs [10]. It uses dual gates, such as the update and reset gates. In the reset gate, it defines the preceding hidden layer (HL); h_{t-1} must be rejected after calculating the novel candidate HL, h_{t-1}. The update gate can be controlled by the number of novel

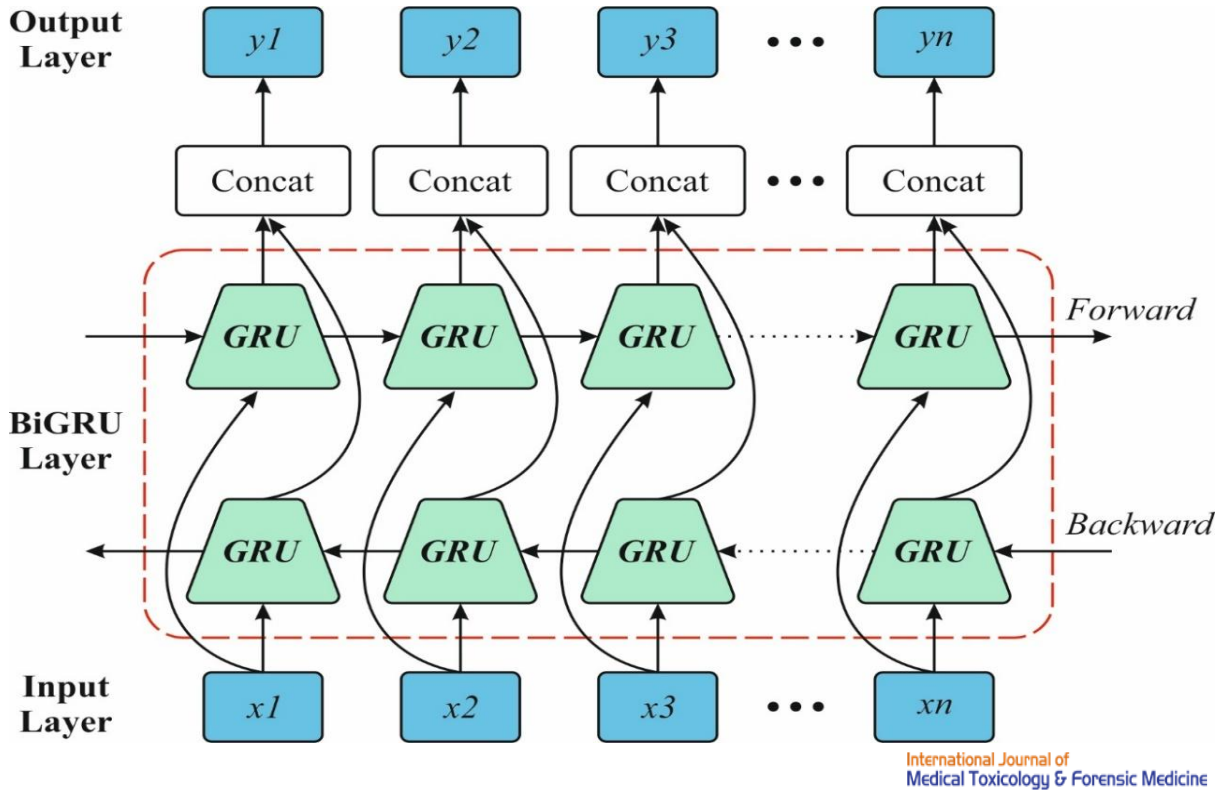

Figure 2. BiGRU architecture.

candidates HLs, $\tilde{h}$, which must be applied to upgrade the present HL.

Hyperparameter Tuning via MOAOA

To optimize the model's performance at a higher level, MOAOA has been employed for hyperparameter tuning, ensuring that the best hyperparameters are selected to boost correctness. The AOA is an original metaheuristic optimization method inspired by the combined quadratic process [11]. The AOA depends on local modification rather than on subtractive or additive methods. These methods are considered to have insignificant dispersion and assist in quick convergence to the goal, thereby advancing the identification of the best solution. The equation is provided as (7)

$$x_{t+1} = \begin{cases} x_{best} - \mathbb{MOP}(t) + \times \left((x_u - x_l) \times \mu + x_l\right), r_3 < 0.5, \\ X_{best} + \mathbb{MOP}(t) \times \left((x_u - x_l) \times \mu + x_l\right), otherwise \end{cases}$$

Now, r_3 denotes randomly generated numbers between (0,1).

The grid model requires the organization of a consistent grid within the main area to calculate possible outcomes, thereby achieving equilibrium among different goals. The leader's selection x_{leader} from the nondominated Pareto set is implemented as,

$$\triangle f_i^n = \frac{\max f_i^n - \min f i^n}{M} \quad (8)$$

Whereas M refers to grid counts, and F_i means the object area value for the ith object. To expand the Pareto Front (PF) range, the process must select non-dominated solutions within the sparse region of the equivalent objective space. The likelihood of the chosen solution in the present region is

$$P_k = e^{\frac{-\beta \cdot m_k}{m_{set}}} \quad (9)$$

Here, k denotes the corner mark of the present region. β refers to the parameter controlling the likelihood of pressure, m_k characterizes solutions in the present region, and m_{set} signifies the total solution counts. In the MOAOA model, the x_{leader} has been taken from the outside record over the roulette-wheel model, as described in Equation (15).

To control the dimensions of the non-dominated set, the method arbitrarily removes solutions when the number of components exceeds the set's current dimensions. The removal procedure specifically selects higher-density solutions to improve similarity within the Pareto set. The reduction in the classification error score has been determined as the FF.

$$fitness(x_i) = ClassifierErrorRate(x_i)$$

$$= \frac{no\ of\ misclassified\ instances}{/Total\ no\ of\ instances} * 100 \quad (10)$$

Results

In this section, the performance of the MOIC-AVCPDL approach is evaluated on two datasets, namely Flickr8k with 8000 images [12] and Flickr30k with 30000 images [13]. Figure 3 indicates samples of original, pre-processed, and feature map images. Figure 4 shows the instances actually generated and the predicted captions.

In Figure 5, the training $accu_y$ and validation $accu_y$ results of MOIC-AVCPDL algorithm at Flickr8k dataset are portrayed. These two data rates are measured in 0 to 25 epochs. It shows improved patterns,

Figure 3. Sample of original, pre-processed, and feature map images.

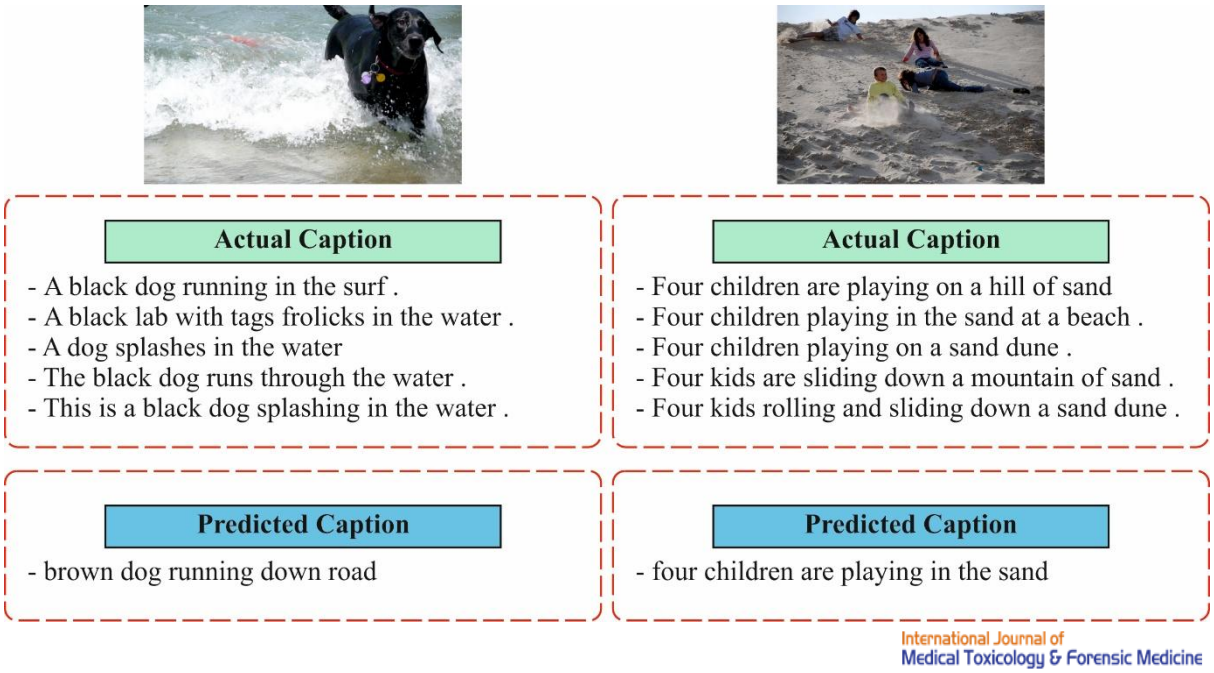

Figure 4. Sample of actual and predicted caption.

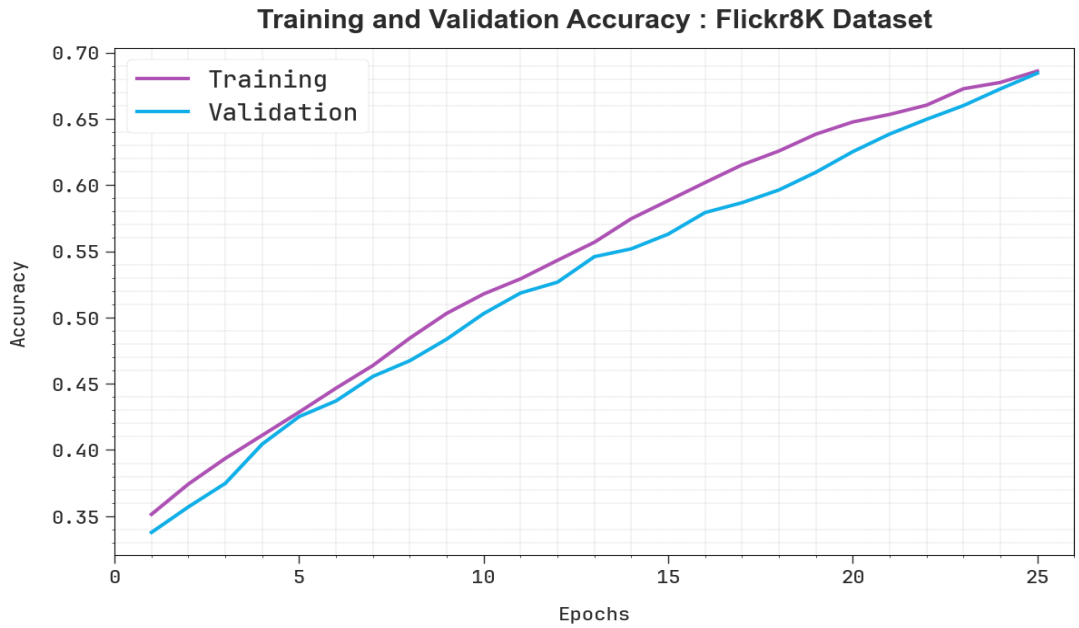

Figure 5. $Accu_y$ curve of MOIC-AVCPDL model on Flickr8k dataset.

indicating the capabilities of the MOIC-AVCPDL method, with enhanced performance in some repetitions. The two methods converge over epochs, specifying minimal overfitting and effective performance.

Discussion

The innovative standard collections, based on sentence-based image search and description, comprise 8,000 individual images, each paired with 5 different captions. The images are selected from six diverse Flickr sets, generally do not feature popular locations or people, and are physically chosen to show a variety of states and scenes.

Table 1 presents a comparison of the FDTLGO-AICSVD model on the Flickr8k dataset against recent techniques across different metrics [14, 15]. Based on BLEU-1, the presented MOIC-AVCPDL model achieved a higher score of 78.08%. But recent methodologies such as NIC, Soft-Attention, Google-NICG, SCA-CNN, Hard-Attention, L-Bi-linear, ResNet-50, AlC-SSAIDL, and MMGAT have achieved lower values of 59.69%, 62.01%, 64.00%, 66.66%, 68.84%, 69.11%, 73.02%, 73.66%, and 76.21%, respectively. Besides, relying on BLEU-4, the presented MOIC-AVCPDL approach has reached a high result of 76.68%, but the recent methods, like NIC, Soft Attention, Google-NICG, SCA-CNN, Hard-Attention, L-Bi-linear, ResNet-50, AlC-SSAIDL and MMGAT have acquired reduced values of 19.58%, 21.42%, 24.69%, 26.06%, 28.18%, 70.86%, 72.49%, 33.22%, and 74.56%, respectively.

Table 1. Comparison study of FDTLGO-AICSVD model at Flickr8k dataset.

	Flickr8K Dataset						
Model	**BLEU-1**	**BLEU-2**	**BLEU-3**	**BLEU-4**	**METEOR**	**ROUGE-L**	**CIDEr**
NIC	59.69	44.46	33.54	19.58	15.98	73.21	32.66
Soft-Attention	62.01	46.37	35.9	21.42	17.7	73.22	35.54
Google-NICG	64	48.06	37.83	24.69	19.34	69.57	38.11
SCA-CNN	66.66	51.34	40.97	26.06	22.94	69.73	39.85
Hard Attention	68.84	53.33	43.01	28.18	25.47	70.07	42.73
L-Bi-linear	69.11	72.25	71.99	70.86	29.37	72.81	69.48
ResNet-50	73.02	72.03	70.51	72.49	32.29	70.4	69.4
AlC-SSAIDL	73.66	57.89	47.56	33.22	29.98	69.58	71.39
MMGAT	76.21	74.18	72.84	74.56	40.87	75.62	73.86
MOIC-AVCPDL	**78.08**	**76.43**	**74.94**	**76.68**	**44.06**	**77.31**	**75.07**

International Journal of
Medical Toxicology & Forensic Medicine

Figure 6 offers the comparative results of the MOIC-AVCPDL algorithm on the Flickr30k dataset on different metrics. The proposed MOIC-AVCPDL approach achieves state-of-the-art results with BLEU-1 at 77.47%, BLEU-2 at 74.11%, BLEU-3 at 72.43%, and BLEU-4 at 74.27%. In the meantime, the recent methodologies, namely Log Bilinear, Google NIC, Soft Attention, Hard Attention, PoS, LRCN, SCA-CNN, AIC-SSAIDL, and GET (w/MAC), have achieved poor performance.

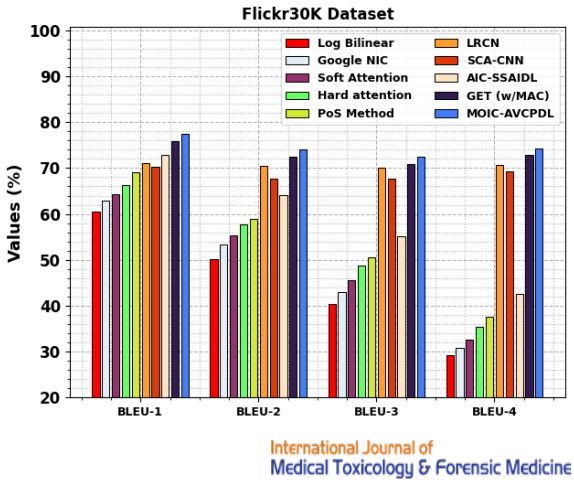

Figure 6. Comparison of the outcome of the MOIC-AVCPDL approach at the Flickr30k database under dissimilar metrics.

Figure 7 exemplifies the comparative results of the MOIC-AVCPDL technique on the Flickr30k dataset at METEOR, ROUGE-L, and CIDEr. The presented MOIC-AVCPDL model achieves a higher value of 50.48%. At the same time, the recent methodologies like Log Bilinear, Google NIC, Soft Attention, Hard Attention, PoS, LRCN, SCA-CNN, AIC-SSAIDL, and GET (w/MAC) have acquired lower values of 23.35%, 24.86%, 27.70%, 29.78%, 30.25%, 30.67%, 28.24%, 36.36%, and 46.75%, respectively. The proposed model achieves significant improvements in image captioning to assist blind persons by effectively integrating image preprocessing, feature extraction, and optimization techniques in real time.

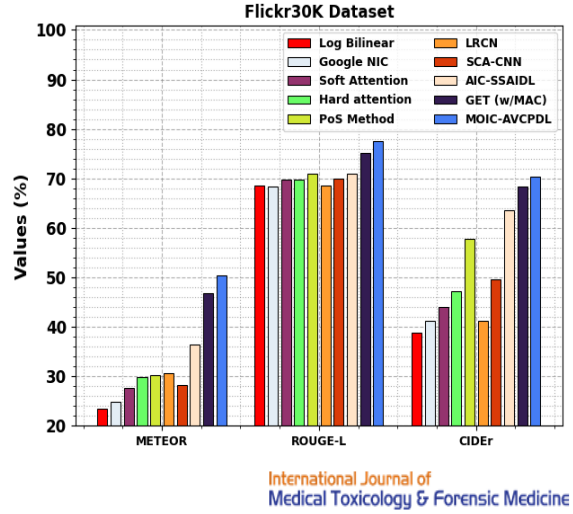

Figure 7. Comparative analysis of MOIC-AVCPDL model on Flickr30k dataset under METEOR, ROUGE-L, and CIDEr.

The presented PCADRS-ODL architecture processes healthcare-related textual information for drug review classification; thus, privacy preservation is a significant consideration for real-world deployment. Although the publicly available dataset used in this study does not contain directly identifiable personal information, healthcare-related textual data collected from real-world healthcare platforms may include sensitive information. Protecting such information is vital to ensure a secure and responsible use of AI-based healthcare applications.

Conclusion

This paper introduces an innovative MOIC-AVCPDL model for efficient image captioning using advanced DL techniques. Initially, WF is employed during preprocessing to improve image quality. Subsequently, the InceptionResNetV2 model is used for feature extraction, followed by the BiGRU for caption prediction. To further improve model performance, the MOAOA is applied for hyperparameter tuning to obtain optimal parameter settings and enhance accuracy. The proposed MOIC-AVCPDL model performs extremely well across multiple performance metrics.

Although the proposed PCADRS-ODL framework demonstrates efficient performance for drug review text classification, it still has several limitations. The current architecture relies on textual information from the Drug Review Dataset and does not integrate dynamic graph structures or evolving relational information; thus, adaptation to evolving data relationships has not been investigated. Moreover, uncertainty estimation, confidence calibration, and prediction reliability analysis are not explicitly evaluated, and the robustness of the techniques under noisy reviews, incomplete textual information, and previously unseen drug-condition combinations requires additional exploration. When the GOA enhances model performance, detailed convergence analysis, computational complexity evaluation, and parameter sensitivity analysis remain significant future considerations. Additionally, privacy preservation is an essential aspect of healthcare-related text processing. The present dataset does not include directly identifiable information; future implementations will investigate privacy-preserving techniques, comprising data anonymization, secure data management, federated learning, and differential privacy, to improve data confidentiality. Future research will also focus on adaptive learning techniques, domain

adaptation methods, and evaluation on larger and more various pharmaceutical datasets to enhance the scalability, reliability, and generalization ability of the proposed approach.

Ethical Approval

Was not applicable.

Acknowledgment

None.

Funding

No external funding was received for this research.

Conflicts of Interest

The authors report that there are no competing interests to declare.

References

[1] Faurina R, Jelita A, Vatresia A, Agustian I. Image captioning to aid blind and visually impaired outdoor navigation. Int J Artif Intell. 2023;12(3):1104-17. [DOI: 10.11591/ijai.v12.i3.pp1104-1117]

[2] Ahsan H, Bhalla N, Bhatt D, Shah K. Multi-modal image captioning for the visually impaired. arXiv [Preprint]. 2021. [DOI: 10.18653/v1/2021.naacl-srw.8]

[3] Ayadi M, Masmoudi N, Almuqren L, Alshahrani HS, Aljohani RO. Designing a novel CNN-LSTM-based model for Arabic handwritten character recognition for the visually impaired person. J Disabil Res. 2025;4(1):20240080. [DOI: 10.57197/JDR-2024-0080]

[4] Safiya KM, Pandian R. A real-time image captioning framework using computer vision to help the visually impaired. Multimed Tools Appl. 2024;83(20):59413-38. [DOI: 10.1007/s11042-023-17849-7]

[5] Valipoor M, de Antonio A, Cabrera J. Analysis and design framework for the development of indoor scene understanding assistive solutions for the person with visual impairment/blindness. Multimed Syst. 2024;30(3):152. [DOI: 10.1007/s00530-024-01350-8]

[6] Hilal AM, Alrowais F, Al-Wesabi FN, Marzouk R. Red Deer Optimization with artificial intelligence enabled image captioning system for visually impaired people. Comput Syst Sci Eng. 2023;46(2). [DOI: 10.32604/csse.2023.035529]

[7] Ayadi M, Masmoudi N, Almuqren L, Aljohani RO, Alshahrani HS. Empowering accessibility in handwritten Arabic text recognition for visually impaired individuals through optimized generative adversarial network (GAN) model. J Disabil Res. 2025;4(1):20240110. [DOI: 10.57197/JDR-2024-0110]

[8] Gupta P, Katal N. Deep learning based automatic image caption generation for visually impaired people. In: Intelligent Systems and Applications in Computer Vision. Boca Raton: CRC Press; 2023. p. 141-57. [DOI: 10.1201/9781003453406-12]

[9] Alruily M, Abd El-Aziz AA, Mostafa AM, Ezz M, Mostafa E, Alsayat A, et al. Ensemble deep learning for Alzheimer's disease diagnosis using MRI: integrating features from VGG16, MobileNet, and InceptionResNetV2 models. PLoS One. 2025;20(4):e0318620. [DOI: 10.1371/journal.pone.0318620]

[10] Mahdi E, Barreiro CM, Cabezas X. A novel hybrid approach using an attention-based Transformer + GRU model for predicting cryptocurrency prices. 2025. [DOI: 10.3390/math13091484]

[11] Chen L, Lin X, Ma L, Wang C. A BiLSTM model enhanced with multi-objective arithmetic optimization for COVID-19 diagnosis from CT images. Sci Rep. 2025;15(1):10841. [DOI: 10.1038/s41598-025-94654-2]

[12] Adityajn105. Flickr8k dataset [Internet]. Kaggle. [Link]

[13] Hsankesara. Flickr image dataset [Internet]. Kaggle. [Link]

[14] Wang B, Wang C, Zhang Q, Su Y, Wang Y, Xu Y. Cross-lingual image caption generation based on visual attention model. IEEE Access. 2020;8:104543-54. [DOI: 10.1109/ACCESS.2020.2999568]

[15] Arasi MA, Alshahrani HM, Alruwais N, Motwakel A, Ahmed NA, Mohamed A. Automated image captioning using sparrow search algorithm with improved deep learning model. IEEE Access.

2023;11:104633-42. [DOI: 10.1109/ACCESS.2023.3317276]