

Original Research Paper

Hybrid Attention-Based Stacked Bi-LSTM Model for Automated MultiImage Captioning

Paspula Ravinder and Saravanan Srinivasan

Department of Computer Science and Engineering, School of Computing Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Chennai, Tamil Nadu, India

Article history

Received: 12-12-2024

Revised: 06-02-2025

Accepted: 19-02-2025

Corresponding Author:

Paspula Ravinder
Ph.D Scholar,
Department of Computer
Science and Engineering,
School of Computing Vel Tech
Rangarajan Dr. Sagunthala
R&D Institute of Science and
Technology, Chennai, Tamil
Nadu, India
Email: vtd703@veltech.edu.in

Abstract: In recent days, the process of medical image captioning is become a prominent field. The distinct characteristics of medical imaging data provide a number of challenges when captioning medical images. Also, the variability in image modalities makes it difficult to generate an effective captioning process. Thus, the proposed study aims to design a novel Multi-image Captioning Hybrid Attention Model to afford effective automated medical image captioning with minimum medical errors. Image acquisition is the initial stage of acquiring input images from the specified dataset. Then, data augmentation is accomplished to maximize the dataset's size. After that, preprocessing is performed to enhance the quality of inputs through Improved Wiener Filtering (IWF), image resizing and color channel conversion. Next, the necessary features are extracted and bounding boxes are generated by utilizing a new Position Attentional YOLOv5 (PA-YOLOV5) approach. Subsequently, the captioning process is performed through the proposed innovative Attention-based Stacked Bi-directional Long-ShortTerm capsule network (A-SBiLSTCN) model. To enhance the efficiency of the proposed model, its hyper-parameters are finetuned by using the Chaotic Flamingo Search Optimization (CFSO) algorithm during the training stage. For experimentation, the Python platform is used, and the simulation is performed using the PEIR dataset. The proposed study outperformed other existing methods in terms of BLEU score (92.87%), METEOR score (88.20%), ROUGE-L score (73.20%), SPICE score (70.76%) and RIBES score (60.40%).

Keywords: Medical Image Captioning, Hybrid Attention, Color Channel, YOLOv5, Optimization, Hyperparameter Tuning, Bleu Score

Introduction

Medical imaging has been a vital tool for doctors to diagnose and treat their patients (Chitteti and Madhavi, 2024; Jaiswal *et al.*, 2024). The technology can be used by medical professionals to construct medical records from CT or X-ray images (Yang *et al.*, 2023). Image captioning can be applied to text-based image extraction, appropriate keyword assignment, human-robot interaction, and assistive technology for the blind and physically impaired.

Convolutional neural network (CNN), template-based, Recurrent Neural Network (RNN), DL-based and other algorithms have been developed for image captioning (Nguyen *et al.*, 2023; Jaruschaimongkol *et al.*, 2024). Auto-image captioning refers to the automated technique of producing human-like descriptions for images. It is a

critical task with broad economic as well as practical ramifications. Auto image captioning is employed in a wide series of industries, including manufacturing, medicine, agriculture, safety, as well as surveillance (Harshitha *et al.*, 2024; Hossen *et al.*, 2024; Yong *et al.*, 2024). It is a very important and demanding task in computer vision.

The goal of conventional object identification and image captioning issues was to identify the objects in the image (Revathi and Kowshalya, 2024; Arasi *et al.*, 2023). In contrast, automatic image captioning is utilized to grasp the complete scene and the relationships between the objects (Thangavel *et al.*, 2023). After fully comprehending, it is critical to provide a human-like account of the occurrence (Deepak *et al.*, 2023; Parvin *et al.*, 2023). Many experiments are being carried out in an attempt to give robots human traits and replace manual

labor with automation and artificial intelligence. The task of producing image captioning results and accuracy on the level with human capabilities has always been incredibly difficult for robots (Li *et al.*, 2023; Chandaran *et al.*, 2023). The number of medical images is always increasing, which adds to physicians' reading and report-writing time. Medical image captioning can aid clinicians by speeding up the reporting process and reducing their workload (Wei *et al.*, 2023). The current methods have produced significant results, resulting in enhanced captioned outputs. Current techniques do not consider how objects and things interact (Derkar *et al.*, 2023; Tiwary and Mahapatra, 2023).

Several deep learning networks are employed to perform the aforementioned important functions. For effectively mining the medical image captions, it is noticed that the encoder-decoder algorithms have attained better outcomes. The CNN layers initially retrieve the image features (Moratelli *et al.*, 2023). After that, the RNN model retrieves shape-related data using the extracted features. Next, LSTM is used to extract the textual data from the images (Mishra *et al.*, 2023; Chen, 2024). End-level tokens are extracted by repeating this process, and encoder-decoder techniques are widely used in medical image captioning (Ravinder and Srinivasan, 2024). Single Long Short-Term Memory (LSTM) is used for the majority of applications. CNN can be used together with other region-presenting techniques, such as RCNN and Faster RCNN (Phueaksri *et al.*, 2023; Mao *et al.*, 2024; Meng *et al.*, 2024), to acquire visual elements and objects for sequential text descriptions. These networks are utilized to develop techniques for automatic image captioning in a variety of fields (Rinaldi *et al.*, 2023; do Carmo Nogueira *et al.*, 2023; Zhai *et al.*, 2023). However, there is still space for development to enable the machine to generate descriptions that are comparable to those written by people (Prudviraj *et al.*, 2023; Verma *et al.*, 2024).

Automatic image captioning through deep learning models is critical in many areas, including remote sensing, social networking systems like producing dynamic websites, Facebook's capacity to immediately infer from images, medical image interpretation, mapping natural language to images and so on. However, image captioning is widely used, and a lot of information is manually identified in images by remote satellites and medical experts. Even for specialists, it is very difficult or impossible to find things in images. The most effective usage of deep learning models is critical for speeding up image interpretation. A successful image captioning system must automatically recognize and label both non-visual and visible objects in an image. In this field, different methodologies are used, most of which are time-consuming or erroneous, as discussed in the literature. However, existing methods are inefficient for captioning multiple images. Due to the reality that obtaining

characteristics across numerous photos is challenging using the current methods.

Limitations of the First Paper

1. Feature Extraction Limitations in YOLOv4

- YOLOv4 is less optimized for smaller objects in medical images
- Struggles with complex spatial relationships in images
- Lower detection accuracy for overlapping regions

2. Sequential Modeling Issues with LSTM:

- LSTM captures sequential dependencies but lacks bidirectional learning, meaning it may not fully utilize contextual information
- Struggles with long-range dependencies in medical image captions
- Higher training time and computational cost

3. Performance Issues:

- The model may have experienced overfitting due to single-directional LSTM
- Lower BLEU, METEOR, and ROUGE-L scores compared to newer architectures
- YOLOv5 has better real-time processing and improved bounding box predictions

4. Optimization Issues:

- Hyperparameter tuning was not optimized effectively in YOLOv4
- The first paper did not incorporate advanced attention mechanisms (e.g., Hybrid Attention in BiLSTM)

Also, overfitting and underfitting are the two main problems experienced by the existing works. Inspired by existing concerns, this study creates a novel DL model for medical image captioned with the goal of improving outcomes and resolving all study-related issues. The graphical abstract of the suggested study is clearly shown in Fig. (1).

The major modules of the suggested effort are listed in detail below:

- To present a novel DL method for medical image captioning to decrease doctors' workload as well as save time and costs
- To introduce a novel Positional Attentional YOLOv5 (PAYOLOV5) model, which can extract the most important feature information and build bound boxes, in order to address the inefficiency issue for learning the features in the previous studies
- To propose a novel Hybrid ASBiLSTCN for effectively captioning multiple images with better performance
- To optimize the network model by finetuning the parameters of the proposed network utilizing the

CFSO process and avoid unwanted errors and underfitting issues in the network

- To implement the suggested approach in the PYTHON platform as well as evaluate the performance measures such as BLEU and METEOR, which show the efficiency of the suggested strategy when related to existing methods

Related Works

Our previous study (Ravinder and Srinivasan, 2024). utilized YOLOv4 for feature extraction and LSTM for sequential modeling in medical image captioning. However, YOLOv4 faced challenges in accurately detecting smaller and overlapping medical objects, leading to lower captioning accuracy. Additionally, LSTM, while effective in sequence learning, struggled with long-range dependencies and lacked bidirectional context learning. These limitations resulted in suboptimal BLEU and ROUGE-L scores. To address these challenges, this study introduces an improved model using YOLOv5 for enhanced feature extraction and BiLSTM for bidirectional sequence modeling, leading to better contextual understanding and performance.

Medical image captioning has undergone significant advancements, transitioning from traditional LSTM-based models to attention-based and transformer-based architectures. Early approaches relied heavily on Long short-term memory (LSTM) networks, where the task was primarily treated as an image-to-text problem. LSTMs, with their ability to handle sequential data, were employed to generate descriptions of medical images (e.g., X-rays and MRIs), but they struggled to capture long-range dependencies in the image's spatial features (Xu *et al.*, 2015). While Bi-LSTM models showed improved performance by leveraging bidirectional context for sequence generation (Vinyals *et al.*, 2015), their performance remained limited by their inability to effectively understand global image features.

With the rise of attention mechanisms, there was a shift toward attention-based models that helped address the limitations of LSTMs. These models enhanced captioning quality by focusing on specific regions of the image, thereby improving accuracy in generating contextually relevant descriptions. Attention-based mechanisms allowed models to generate more accurate and targeted captions by selectively focusing on parts of an image that were most relevant to the caption (Anderson *et al.*, 2018). However, despite their ability to focus on important features, these models still faced challenges in handling complex medical terminologies and multimodal data.

In recent years, the field has seen the emergence of transformer-based architectures such as Vision Transformer (ViT) and hybrid models like Swin Transformer, which combine CNNs and transformers for better feature extraction and caption generation (Dosovitskiy *et al.*, 2016). Transformer models have revolutionized medical image captioning by allowing models to capture both local and global image features through self-attention mechanisms, resulting in more precise and clinically relevant descriptions. Models like ViT + GPT-2 (Radford *et al.*, 2019) and BioViL-T (Li *et al.*, 2019) have demonstrated significant improvements, especially in handling the complexities of medical images and medical terminology. These architectures, despite their effectiveness, require large datasets and computational resources for training, often making them challenging to deploy in resource-constrained environments.

The evolution from LSTM-based to attention-based and transformer-based architectures marks a key progression in the accuracy and reliability of medical image captioning. With advancements in Coupled Feature Selection Optimization (CFSO) and the integration of domain-specific knowledge, the latest models continue to improve caption generation for medical image analysis, offering significant potential for clinical applications such as radiology, pathology and surgery.

Djamila-Romaissa *et al.* (2021) used a CNN-based multi-label classification method for identification as well as an attention-based encoder-decoder strategy for caption prediction in the image CLEF 2021 medical task. Both methods used Centre cropping-based image enhancement to increase the training sample size, while transfer learning was used to extract essential features from actual radiological images. Unfortunately, this technique needs little data, affecting the system's accuracy. This method would generate a lot of false positives or images that would be incorrectly identified if it were limited to a lower number of classes.

Chen *et al.* (2022) presented Visual-GPT, a data-efficient image captioning algorithm that combines

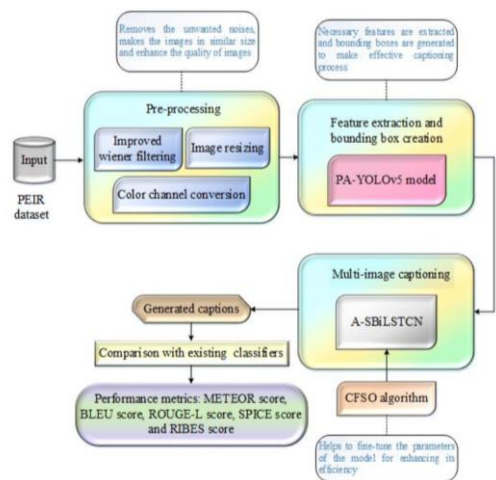


Fig. 1: Graphical abstract of the proposed study

linguistic data from a pretrained language system. To bridge the semantic gap between different methods, the author developed encoder-decoder attention methods that use an unbalanced corrected gated function. Nonetheless, linguistic data from pre-trained algorithms is most beneficial when training data is scarce and fails to cover the whole lexicon. As the in-domain training information is more, the difference between baseline models and Visual-GPT rapidly disappears.

Zeng *et al.* (2020) suggested the detection technique to create ultrasound image captions. This method mainly fixes the problem of multiple organs interfering with each other while also gathering more accurate data on the focal region. The region recognition methodology is initially utilized to identify focus areas, and encoding vectors are then discovered. This study used LSTM to decode the encoding vectors and produced explanation texts for ultrasound images that identify target areas' content data. However, the accuracy and pixel quality obtained in this research were highly degraded.

Zhou *et al.* (2019) utilized an image captioning strategy combined with object features to address a problem with the present widely used DL approach for producing phrases. The integration of an attention mechanism and an object recognition network reduces the number of errors in generated descriptions, efficiently reduces the object category and improves the quality of created sentences. The image's data on the object region, number and category is acquired. However, this model's captioning is inaccurate and cannot be used for human activity. The algorithm needs time to process the captions.

Singh *et al.* (2022) suggested an updated show, Attend and Tell (ATM) paradigm for image captioning. ATM is an optical concentration approach that relies on encoder-decoder design. The fundamental parameters were optimized with an ATM based on the Strength Pareto Evolutionary Algorithm-II (SPEA-II). SPEA-II was employed in this previous study to finetune the initial characteristics of SPEA-II-based ATM. Numerous trials have shown that SPEA-II-based ATM surpasses current medical image captioning methods. The model parameters are only finetuned using SPEA-II. However, the ATM strategy fails since there is no viable metaheuristic procedure.

Lin *et al.* (2023) created a multi-class classification model to caption skin medical images. This existing study included four learning models: Discriminator, autoencoder, multi-class classification model and Siamese network. Initially, the discriminator segments the background skin regions and extracts the required features using the auto-encoder approach. The features produced by the discriminant, as well as the autoencoder, are then

merged to produce the keyword labels. Finally, text similarity is determined to create an effective captioning process. However, computational cost is enhanced and is considered a limitation of this existing study.

Sharma and Srivastava (2023) used multilevel attention and relation networks to caption the input images. This existing study used a Local Relation Network (LRN) to analyze the correlation between features. Furthermore, the developed model captured the necessary features associated with the image region. In addition, attention-based LSTM is employed to obtain relevant contextual information on deep visual and spatial features. Based on the recorded attributes, the developed model encoded the images and extracted crucial cues for improved caption generation. However, the developed model's learning ability is limited due to its inefficiency.

Kong *et al.* (2024) combined the CNN classification method with GPT-2 to produce text for Intracerebral hemorrhage (ICH) in a sequential manner. In this example, CNN is in charge of determining the availability of ICH from the given CT images as well as collecting feature information from 3D ICH CT images.

These feature vectors are trained to create captions for each CT image and supplied with the text in GPT-2. The pre-trained CNN classifiers used in this investigation are ResNet-50V2, DenseNet-121, VGG-16 and VGG-19. The simulation outcomes show that the developed method achieved better results; however, this study faced a complex training process. Table (1) analyses details of certain state-of-the-art approaches associated with image captioning.

Ravinder and Srinivasan (2024) proposed an automated medical image captioning system integrating a soft attention-based LSTM model with the YOLOv4 algorithm for efficient object detection and caption generation (Ravinder and Srinivasan, 2024).

Problem Statement

The rapidly developing field of artificial intelligence is driving acceptance in the automatic production of visual descriptions. The purpose of image captioning is to use the data controlled in an image to construct phrases that are both linguistically and semantically reliable. Various techniques are used to reduce errors, save time and accomplish many other goals. Techniques like CNN, RNN, LSTM and Fast RCNN are well-known deep learning models. Large parameters probably lead to overfitting, but these strategies have many drawbacks, including an increasing number of variables to learn with increasing width and depth.

Table 1: Details of some state-of-the-art approaches

Reference	Technique name	Advantages	Disadvantages	Performance measures
Xu <i>et al.</i> (2015)	CNN-based multi-label classification model	The obtained measure score of 14.30%	Achieved higher false positive rates	F_Measure, recall, precision, BLEU, Brevity Penalty
Vinyals <i>et al.</i> (2015)	VisualGPT	VisualGPT exceeds the finest model by up to 10.0%	The gap among baseline methods, as well as VisualGPT, progressively disappears as in-domain training data rises	-
Anderson <i>et al.</i> (2018)	Create captions for ultrasound images based on 360-degree region recognition	Attained a mAP value of 75.4% on focus area recognition, reduced running time as well as parameters	The faster this algorithm operates, the more accuracy and pixel quality will decrease	TP, FP and FN
Dosovitskiy <i>et al.</i> (2016)	Image captioning model fused with object features	Efficiently decreases the errors of object group and improves the quality of generated sentences	The model cannot be used for human action, and the processing time is high	Accuracy and recall rate
Radford <i>et al.</i> (2019)	Show, Attend and tell model (ATM)	ATM can recognize captions with good performance	ATM agonizes from Hyper-parameter tuning issues	Visual Analysis, Quantitative Analysis
Li <i>et al.</i> (2019)	Multi-class classification model	Efficient features are extracted	Higher computational cost	Accuracy, binary accuracy, MSE, MAE
Djamila-Romaissa <i>et al.</i> (2021)	LRN, LSTM	The lightweight model completes the task with reduced processing time	Reduced learning ability	BLEU score, METEOR, ROUGE, SPICE
Chen <i>et al.</i> (2022)	CNN	Requires less memory	Faced complex training process	Precision, recall, accuracy, F1-score

Furthermore, computational resources are consumed by the straightforward stacking of convolutional layers. The CNN approach may lack crucial image data during the feature extraction stage. The LSTM cannot be trained in parallel, and it requires a large dataset and a lot of memory. RNN training should be difficult because it consumes less computational time. It has problems with explosions and gradient disappearance. In addition, by surveying several existing works, training complexity, higher computational cost, and reduced learning ability are the major limitations of captioning. To solve these limitations, it is crucial to create an effective model; this study introduces a novel strategy to lessen these problems and achieve better performance.

Materials and Methods

The materials utilized in this study include the PEIR dataset (<https://peir.path.uab.edu/library/index.php?category/106>) for medical image captioning, along with various preprocessing and deep learning methodologies. The experimental setup involved using Python for implementation, an Intel Core i5 processor with 8GB RAM, and the YOLOv5-based PA-YOLOV5 feature extraction model. The study also incorporated the A-SBiLSTCN model for caption generation, which was optimized using the CFSO algorithm. The primary

equipment used included high-performance computing resources and the TensorFlow deep learning framework for model training and evaluation.

The suggested project intends to create an efficient automatic medical image labeling method to reduce the doctor's burden and thereby save money as well as time. The suggested Model system involves four essential stages: Image acquisition, preprocessing, feature extraction-bounding box creation and multi-image captioning. At first, the image collection step collects input images from the specified dataset and performs data augmentation to raise the dataset size. The raw input images have greater noise, which can degrade system efficiency. The proposed approach provides an effective preprocessing stage by improving wiener filtering, image scaling and color channel conversion. These preprocessed images are given into the PA-YOLOv5 algorithm, which extracts key features and generates boundary boxes. The created bounding box localizes the target, hence improving detection accuracy. The proposed A-SBiL STCN then completes the labeling procedure. Here, the loss in the proposed captioning model is minimized via the CFSO algorithm. This optimization approach optimizes the proposed model's hyper-parameters (learning rate, maximum epochs, input size, hidden and output layers) to improve detection performance.

Preprocessing Stage

Initially, the preprocessing stage is carried out to improve the medical captioning presentation of the network approach. Preprocessing involves three key operations: Image denoising, rescaling, and color channel conversion. After using the IWF technique to eliminate noise from the raw dataset, rescaling and conversion procedures are carried out.

Noise Removal Utilizing the IWF Method

The pictures in the raw dataset are rather noisy; thus, minimizing noise is crucial for blurry images since noise from the background is not removed. An IWF method is suggested in this study to address this concern. The result of a noised image $g_{(a,b)}$ is shown below:

$$g_{(a,b)} = f_{(a,b)} \times u_{(a,b)} + n_{(a,b)} \quad (1)$$

Here, $f_{(a,b)}$ denotes the obtained image, $u_{(a,b)}$ signifies degradation function, and $n_{(a,b)}$ specifies noise. The de-noised ending result image $h_{(a,b)}$ from conservative WF is shown below:

$$h_{(a,b)} = WF(g_{(a,b)}) \quad (2)$$

The noised images are fed into the traditional noise reduction filter to generate a de-noised image. The Weiner and median filters act as noise reduction filters that contain the nonlinear spatial domains to produce the de-noised images effectively. Some major steps are followed to enhance the image quality. Initially, the mask matrix size $x' y$ is determined for the utilized spatial noise reduction filter. In the following step, the mask matrix is compared to the new pixel value to determine the difference between the new pixels and the mask pixel size. A median filter is offered, which converts the image pixel value to the median pixel value using the mask pixel value.

As a result, the noise is eliminated, yet the image's original image remains intact. The WF method contains both variances as improving captioning performance. Standard Weiner filtering (Hepsiba and Justin, 2022) results in a well as mean pixel values in the size $x' y$ mask matrix (Zhou *et al.*, 2019; Singh *et al.*, 2022; Lin *et al.*, 2023; Sharma and Srivastava, 2023; Kong *et al.*, 2024; Hepsiba and Justin, 2022; Arazm *et al.*, 2017):

$$\mu = \frac{1}{XY} \sum_{x,y \in \beta} p(x,y) \quad (3)$$

$$\sigma^2 = \frac{1}{XY} \sum_{x,y \in \beta} p(x,y)^2 - \mu^2 \quad (4)$$

where, μ signifies mean, σ^2 signifies modification of Gaussian noise, x, y signifies contiguous area β in the mask, $p(x, y)$ specifies pixels in the area β .

$$R_{(x,y)} = \mu + \frac{\sigma^2 - \vartheta^2}{\sigma^2} * (p(x,y) - \mu) \quad (5)$$

Here, ϑ^2 represents the noise variance obtained from the WF technique. However, the variance between the mask and de-noised images is high. This is due to a lack of noise removal in the background region, which can be remedied by incorporating a median filter into the WF method. The main advantage of WF is that it preserves the edge and modifies the WF pixel values to the median pixel values. The mean obtained from the WF approach is converted into the mean of the MF technique, which may be mathematically expressed as follows:

$$\tilde{R}_{(x,y)} = \tilde{\mu} + \frac{\sigma^2 - \vartheta^2}{\sigma^2} * (p(x,y) - \tilde{\mu}) \quad (6)$$

Here, $\tilde{\mu}$ represents the modified mean $\tilde{R}_{(x,y)}$ denotes the noise-removed pixel obtained from the MF technique.

After de-noising the image, a rescaling operation is performed to reduce unwanted overfitting in the network model. Each raw image has a varied initial size, so all medical images are resized to an effective, ideal size using a rescaling technique.

Colour Channel Conversion

The sequence of pixels in every input picture is determined to evaluate the color channels. The suggested research developed a color vector for color channel evaluation that included a Hue Saturation Value (HSV) color histogram. The HSV of every pixel in the supplied input image is translated to the RGB explanation utilizing subsequent sources (Vo and Verma, 2016):

$$H = \cos^{-1} \frac{1/2[(R-G)(R-B)]}{\sqrt{(R-G)^2 + ((R-B)(G-B))}} \quad (7)$$

$$S = 1 - \frac{3[\min(R,G,B)]}{R+G+B} \quad (8)$$

$$V = \left(\frac{R+G+B}{3}\right) \quad (9)$$

This process requires removing the border of each input image. The color channel is chosen to represent the image's improved quality. The preprocessed input images are supplied into the feature extraction phase, which excerpts useful information from the medical image.

Feature Extraction Using Position Attentional Yolov5

Feature extraction is an important step in extracting significant characteristics from medical images, and it is primarily responsible for enhancing performance. Numerous recognized methods have been used to remove valuable elements from photos. However, these approaches are very consuming and have overfitting concerns. To address these issues, the You Only Live Once (YOLO) version 5 approach is used in this study. The YOLOv5 (Pham *et al.*, 2023) is one of the most effective object detectors among existing versions,

including the YOLOv1, YOLOv2, YOLOv3, as well as YOLOv4 models. The suggested YOLOv5 model solves the correlation concern by combining all items into a single level based on the preceding forms. Because Pytorch is utilized instead of darkNet in this approach, it differs significantly from prior versions.

The main backbone of the YOLOv5 is the CSPDarkNet53. The proposed study implements a position attention technique to improve YOLOv5's feature learning capabilities. This approach allows a wider range of features to be encoded into local features, which improves representation capability. Furthermore, the position attention block focuses on the spatial correlations observed in the feature maps, which improves the YOLOv5 model's performance. The suggested network model consists of three layers: Head, neck, and backbone. Figure (2) demonstrates the YOLOv5 Architecture.

Convolution and Pooling Layer

The convolutional layer extracts key information from the input pixels, resulting in vectors (features). The pooling layer in the network model significantly contributes to lowering the dimensionality of the feature maps.

Backbone Layer

The proposed architecture's backbone is made up of a cross-stage partial network (CSPNet) as well as a darkNet, which combine to form CSPDarkNet. The

CSPdarkNet enhances the convolutional layer to learn the features effectively. Also, this CSPdarkNet eliminates the replication of gradient insufficiency within the network model. To enhance the overall performance, the complexity of a network is reduced during each training.

Furthermore, the Spatial Pyramid Pooling (SPP) block is implemented to improve the accessible field and aid in the extraction of critical elements from the backbone.

The YOLOv5 utilizes the image as input, extracting features using convolutional layers. In addition, a max pooling operation is utilized to obtain the feature sets effectively. The feature sets are pyramidal in shape, measuring width, height and length. Using larger backbones decreases gradient insufficiency, which is a major contributor to boosting captioning performance speed and decreasing complex parameters.

Neck Layer

A Path Aggregation Network (PANet) is used to ensure proper data flow, and a cat serves as the network model's neck. This PANet employs the novel feature-based pyramid network with many top and bottom layers. This further eliminates unwanted low-level elements from the network model. Furthermore, YOLOv5's PANet aids in localizing the high-level features without using additional memory. The image is initially analyzed by CSPdarkNet to obtain features, which are then given into PANET for fusion. Finally, output is obtained in the YOLO layer.

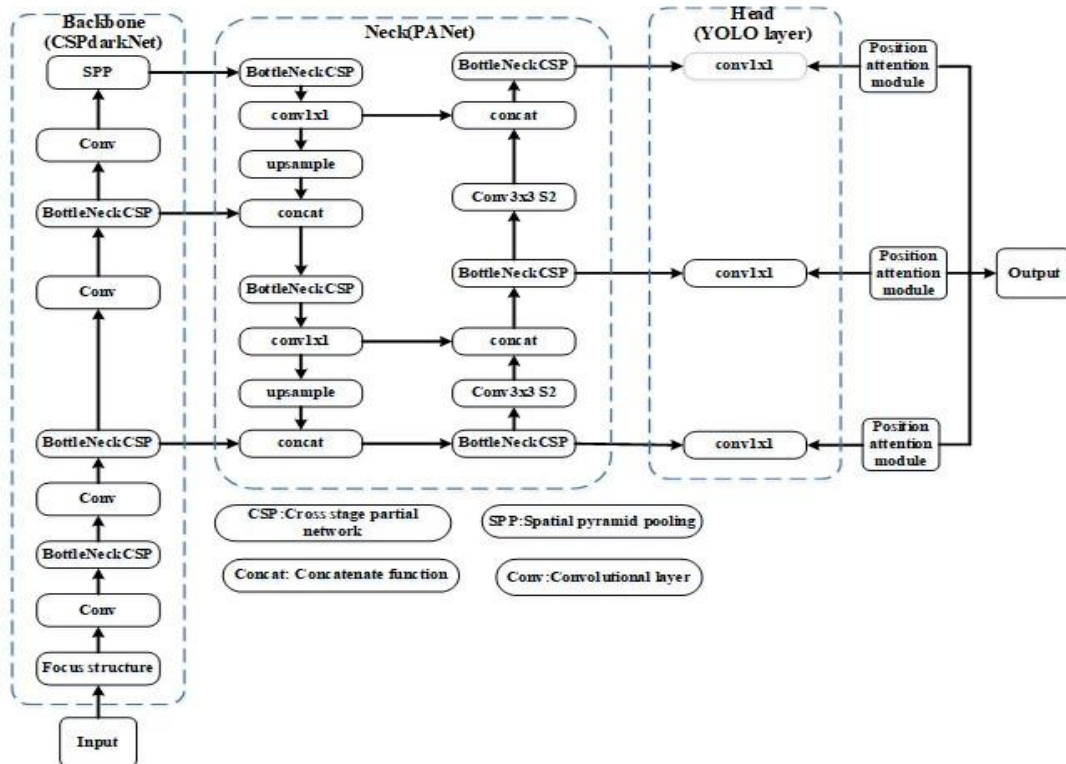


Fig. 2: PA-YOLOv5 architecture

Head Layer

The YOLOv5 uses the same head portion as the YOLOv4 and YOLOv3 models, resulting in multilevel features for improved performance. The head part is primarily responsible for detecting both small and large items. Furthermore, the adaptive pooling layer combines all features from each sub-layer. The primary distinction between YOLOv3, YOLOv4, and YOLOv5 is that YOLOv5 employs Darknet53 as its backbone. YOLOv4 employs the CSPdarknet53 as its backbone, whereas YOLOv5 employs the focus-based CSPDarknet53. The focusing layer is a new layer in YOLOv5 that can accelerate the training process by successfully optimizing both forward and backward propagation. The neck in the YOLOv5 helps the generation of pyramidal features and the effective extraction of multilevel features in various sizes and dimensions. Learners can acquire identical items in a range of scales and sizes because the qualities of the object can be readily generalized across different dimensions.

The bottom-up approach relies on a feed-forward operation in the convolutional layer to build a pyramidal-based bounding box at each level. YOLOv5's result layer serves as the feature reference set, producing high-level features from the semantic layer. To accurately transmit vectors based on bounding box coordinates, confidence value and caption probabilities, the head, known as the YOLO layer, anticipates the objects based on their position.

Position Attention Layer

The position attention layer aids in modeling the rich feature informative relationship over local attributes. In this, the local features are provided as the input of convolutional layers for correspondingly generating feature maps A and B , where the local feature is represented as $L \in R^{B \times M \times N}$ and $\{A, B\}^{R^{B \times M \times N}}$ which is replaced as $R^{B \times P}$. Here, $P = M \times N$ is considered as the pixel count. Then, a matrix multiplication among transpose of B as well as A , which utilizes the softmax layer, is used for determining the spatial attention map $S \in R^{P \times P}$:

$$s_{ij} = \frac{\exp(A_j \cdot B_i)}{\sum_{j=1}^P \exp(A_j \cdot B_i)} \quad (10)$$

where, s_{ij} determines the impact of the position of i and j , the most closely related feature maps of the two positions contribute to stronger relationships between the feature sets. Also, feature L is fed into the convolutional layer to attain a new feature map $F \in R^{B \times M \times N}$ and reshape ^{e^{it}} $R^{B \times P}$. Then, matrix multiplication is performed among F as well as the transpose of S , and the result is redesigned into $R^{B \times M \times N}$. The

scale parameter β is multiplied with the result and enables an element-wise sum process with the attributes L for attaining the final output $H \in R^{B \times M \times N}$ as:

$$H_i = \beta \sum_{j=1}^P (s_{ij} F_j) + L_i \quad (11)$$

where, β is considered as zero and allocates more weight by learning the information in a gradual manner, the global informative features are selectively aggregated through the spatial attention map. Thus, the proposed PA-YOLOV5 method extracts the most significant features from the given inputs.

Image Captioning Using A-SbiLstcn

The extracted features are finally given to the proposed Hybrid A-SBiLSTCN technique to caption the obtained image captioning effectively. The traditional BiLSTM approach (Li *et al.*, 2022) requires a significant amount of training time and is often prone to overfitting. To solve this issue, an attention mechanism is integrated with the BiLSTM technique, resulting in much-improved model performance. The BiLSTM is a kind of RNN that can help decipher the intricate structures of sequential data. Also, the proposed study used a stacked connection of BiLSTM networks to produce image captioning outcomes. In contrast to the basic BiLSTM, the stacked BiLSTM combines multiple layers to create an efficient hierarchical representation-learning model. Capsule networks capture the spatial relationships between features, which improves the model's capacity to identify objects and attributes. Additionally, the stacked connection offers greater model capacity, which is recommended when selecting stacked BiLSTCN for image captioning.

Layers and Gate Operations

The Bi-LSTM model has three layers: Input, hidden, and output. The input layer is mainly accountable for providing input in both left-to-right as well as right-to-left directions. The hidden layers of this system's network are situated among input as well as output layers. The output layer shows the results of training the network system.

Additionally, Bi-LSTM has three gates: Input, output, and forget. The input gate controls the number of data in the existing cell state. The forget gate controls data flow by deleting undesirable past-state data from the memory cell. The output gate regulates data delivered to the following time step.

Bi-LSTM uses forward as well as backward propagation procedures to effectively reduce difficulties that cause the network model to overfit. Figure (3) controls the architecture of the proposed ASBi-LSTCN model.

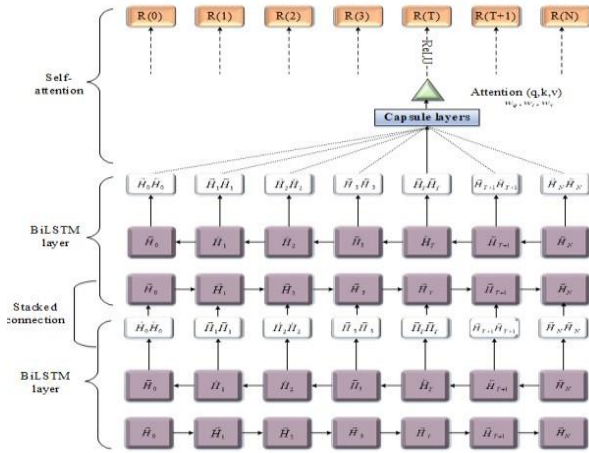


Fig. 3: Architecture of A-SBi-LSTCN model

The consecutive documents sequence can be stated as $p_{task}=[p_0, p_1, p_2, \dots, p_t, \dots, p_N]$ where p_t represents procedures of t steps. Then, the model assesses hidden cell results in H_{t-1} as well, as the outcome is signified O_{task} , as well as output o_t , is resolute using H_t (Anderson *et al.*, 2018):

$$a_t = \tanh(ua) \quad (12)$$

$$H_t = wH_{t-1} + a_t + bias_H \quad (13)$$

$$O_t = vH_t + Bias_O \quad (14)$$

where, u signifies the weight of the input layer, w specifies the weight of the hidden layer, v specifies the weight of the output layer, and u , as well as v , is given to each step correspondingly. With the addition of an attention layer, the suggested A-SBiLSTCN model effectively eliminates vanishing gradient difficulties while also improving every cell state function. Some of the cell state operations are forget gate F , input gate A , as well as cell state C (Anderson *et al.*, 2018):

$$F_t = (w_{FH}H_{t-1} + w_{Fp}p_t + bias_F)\sigma \quad (15)$$

$$A_t = (w_{aH}H_{t-1} + w_{ap}p_t + bias_A)\sigma \quad (16)$$

$$C_t = F_t C_{t-1} + A_t \tanh(w_{cH}H_{t-1} + w_{cp}p_t + bias_C)\sigma \quad (17)$$

$$Z_t = (w_{ZH}H_{t-1} + w_{Zp}p_t + bias_Z)\sigma \quad (18)$$

$$H_t = Z_t \tanh(C) \quad (19)$$

where, F_t specifies the weight of the forget gate σ reveals sigmoid function, which limits the F_t under range $[0,1]$. If the range is 0, preceding data is lost; if it is 1, preceding

data is retained. Furthermore, A_t signifies input cell weight, and C_t denotes current data at the input gate. The result of the Bi-LSTM cell H_t is found by regularizing the tanh function as well as the update gate Z for a responsible amount of information. In order to further enhance the performance of BiLSTM, the suggested study considers a stacked BiLSTM structure. The network model is able to capture a hierarchical representation of the inputs by stacking multiple BiLSTM layers. Each layer helps to extract intricate patterns and relationships from the input samples and can achieve a variety of informative features. Furthermore, the stacked BiLSTM can represent long-range dependencies, increasing the network's efficiency for captioning images effectively. In this study, a stacked BiLSTM network is utilized, and the output is no_t . The mathematical representation of stacked BiLSTM is given as (Cai *et al.*, 2019):

$$H_t = w_{HH} \cdot \vec{H}_t + w_{HH} \cdot \vec{H}_t + bias_H \quad (20)$$

By processing with stacked BiLSTM, the hidden state matrixes denote Q_A and Q_B are obtained as:

$$\begin{aligned} H_t &= SBiLSTM(H_{t-1}^A, H_{t+1}^A, A_t), & H_0^A &= 0, \\ H_t^B &= SBiLSTM(H_{t-1}^B, H_{t+1}^B, B_t), & H_0^B &= H_m^A, \\ H_A &= [H_1^A, H_2^A, \dots, H_m^A] \in V^{d \times m}, \\ H_B &= [H_1^B, H_2^B, \dots, H_m^B] \in V^{d \times n} \end{aligned} \quad (21)$$

where, d represents the hidden state dimension, the capsule network is an advanced neural network structure produced to solve spatial context problems (Zhai and Zhao, 2024). This is accomplished by encoding the parameters' spatial correlations in vectors, allowing the neural network to learn both feature classification and distances from recognized features. CapsNet comprises three layers: Conv1d, Digitcaps and PrimaryCaps. The capsule network essentially replaces the traditional scalar value with a transformation vector weight, which is defined formally in Eq. (22), by performing an affine alternation:

$$v_{ij} = W_{ij} v_i \quad (22)$$

where, \hat{v}_{ij} denotes prediction vector, W_{ij} indicates weighting metrics, and v_i represents output of capsule I vector. By capturing the spatial surroundings, this transformation vector covers the gaps in the standard weight process. The sum of weight vectors is determined in Eq. (23):

$$s_j = \sum_i c_{ij} \hat{v}_{j,i} \quad (23)$$

Thus, c_{ij} represents the coupling coefficient in order to maintain the matrix information provided in the capsule,

an original kind of activation known as the nonlinear pressing function, which is given in Eq. (24):

$$u_j = \frac{\|s_j\|^2}{1+\|s_j\|^2} \times \frac{s_j}{\|s_j\|} \quad (24)$$

Hence, the output u_j was evaluated utilizing a squash function; s_j represents a squashing function. This function produces a squashing output from 0-1 while retaining length as well as spatial data of the provided values, which is the inverse of traditional activation methods.

The attention function can be signified as a query vector (q), key vector (k), as well as value vector (v):

$$Attention(q, k, v) = soft \max\left(\frac{qk^T}{\sqrt{D_k}}\right)v \quad (25)$$

where, D_k employs dimensions into the dot product of both query q and key k , with the use of an attention apparatus, q , k , and v are determined. As a result, query $q = [\vec{H}_t, \vec{H}_t]w_q$, $k = [\vec{H}_t, \vec{H}_t]w_k$, $v = [\vec{H}_t, \vec{H}_t]w_v$. Finally, the forward and backward BiLSTCN results are merged with the attention result and fed into the dense layer to get the final image.

Loss Function

Nevertheless, some of the important features and potential losses are obscured by the network model. Due to this loss, the error gets minimized. The loss function of the suggested network can be represented as:

$$Loss(L) = \max(RMSE) \quad (26)$$

To overcome this issue, CFSO is proposed to tune the network.

Network Model Optimization Using the CFSO Technique

Losses can be reduced by using cutting-edge optimization approaches while the efficiency of the suggested network model improves. However, this method requires too many repetitions to maximize the network approach. To address this concern, a unique CFSO optimization approach is suggested to optimize network characteristics, resulting in enhanced efficiency. A proposed CFSO strategy's primary advantage (Zhiheng and Jianhua, 2021) is that it requires fewer iterations and avoids settling into local optima. Flamingos' key behavioral qualities are foraging and migration. Most flamingo species dwell in areas with an abundance of food. Following a period of intense hunting, flamingo populations relocate anytime there is not enough food in the area to maintain the colony. This CFSO approach finetunes the learning rate of the proposed A-SBi-LSTCN

model to make a better image captioning process. This is because finetuning the learning rate parameter is critical for improving the performance of the created model during the training stage. The learning rate is the step size employed in the weight update of the proposed A-SBi-LSTCN model, as defined by the gradient of the loss function. By altering the learning rate, the proposed study can speed up the training process. Furthermore, finetuning the learning rate of the suggested model can improve training stability. Thus, the proposed study selects the learning rate parameter for finetuning.

Initialize the population to T_p ; here, $Iter_m$ is the maximum number of iterations, and pdf denotes the initial position of traveling search agents.

$mp_s = rand[0,1]$ 'p' $(1-mp_f)$ indicates the amount of foraging search agents in the j^{th} iteration. $mp_q = mp_p$ is the amount of traveling search agents in the initial repetition. The second iteration $mp_d = p - mp_q - mp_s$ is the amount of traveling search agents. Previous search agents mp_h , as well as mp_d with low as well as high fitness, are measured traveling search agents, correspondingly.

Table (2) will clearly display the range of values explored for each hyper-parameter and the final converged values after optimization using CFSO. It will provide a detailed insight into the optimization process that has been applied in your model.

Foraging Behaviour

Communicative Behaviour

The search agent that satisfies most hyper-parameters calls other agents to explore the specific position. The population with the greatest number of parameters can be found where the majority of search agents are available. The supposition is made that the search agent with the greatest parameters in the k^{th} dimension is z_{fk} .

Scanning Behaviour

When identifying hyper-parameters, search agents employ particular scanning techniques on the prey. The search agents examine the area's available resources while foraging. To select the best hyper-parameters, a search agent must first extensively scan the area. The position of the search agent in the k^{th} dimension of the search agent community is z_{jk} .

Table 2: Hyper-parameters optimized using CFSO

Hyperparameter	The range of values explored	Final converged value
Learning rate	0.0001-0.01	0.001
Momentum	0.5-0.99	0.85
Batch size	16, 32, 64	32
Weight decay	0.0001-0.01	0.0005
IoU threshold	0.5-0.75	0.7
Confidence threshold	0.2-0.9	0.5

Consequently, when performing foraging, the search agent's scan length reached the maximum distance, which can be calculated as $|N_1' z_{fk} + \gamma_2' z_{jk}|$; here, γ_2 is a chance value of -1 or 1. Where N_1 is a random value with a typical normal distribution, its alteration curve evenly replicates the alteration of a search agent's recognition range as $N_2 \times |N_1 \times z_{fk} + \gamma_2' z_{jk}|$; here, N_2 is a random variable.

Bipedal Mobile Behaviour

While hunting, search agents go on the path of locations with a significant number of restrictions and examinations. Given that the population's center of abundance for parameters is z_{fk} , distance traveled can be conveyed as $\gamma_1' z_{fk}$ here γ_1 is a random value among 1 and 1:

$$f_{jk}^l = \gamma_1 \times z_{fk}^l + N_2 \times |N_1 \times z_{fk}^l + \gamma_2 \times z_{jk}^l| \quad (27)$$

The equation for changing search agents' hunting behavior positions is:

$$z_{jk}^{l+1} = (z_{jk}^l + \gamma_1 \times z_{fk}^l + N_2 \times |N_1 \times z_{fk}^l + \gamma_2 \times z_{jk}^l|) / M \quad (28)$$

Equation (28) z_{jk}^{l+1} shows the position of the j^{th} search agent in the population's k^{th} dimension in the $(l+1)$ iteration, as well as z_{jk}^l , specifies the location of the j^{th} search agent in the population's k^{th} dimension in the l^{th} iteration. The search agent with the highest fitness in the population during iteration l is characterized by the coordinates z_{fk}^l in the k^{th} dimension. A diffusion factor, or $M = M(n)$, is a chance amount chosen at random from the chi-square distribution with n freedom degrees. The random integers $N_1 = E(0, 1)$ and $N_2 = E(0, 1)$ have a typical normal distribution of $\phi \gamma_1$ and γ_2 , which are measured by -1 or 1.

Migration Behaviour

The search agent species that travels to the next collection of restrictions is sparse in the current foraging zone, which is abundant with parameters. The equation for the search agent population's migration is as follows, presuming that z_{fk} signifies the location of most parameters area in the k^{th} dimension:

$$z_{jk}^{l+1} = z_{jk}^l + \eta \times (z_{fk}^l - z_{jk}^l) \quad (29)$$

Equation (29) z_{jk}^{l+1} denotes the position of the j^{th} search agent in the k^{th} community dimension in the l iteration z_{jk}^{l+1} and position of the i^{th} search agent in the k^{th} dimensions of the community in the $l+1$ iteration. The position of the search agent with the highest fitness in the community during iteration l is signified by z_{fk}^l . The search agent migration search space is expanded

utilizing Gaussian random $\eta = E(0, n)$. In this proposed study, circle chaotic map representation helps to accelerate the FSO method's convergence to the optimum solutions by identifying suitable search field locations. The circle chaotic map is mathematically expressed below the equation:

$$y_{k+1} = y_k + b - \left(\frac{p}{2\pi}\right) \sin(2\pi y_k) \quad (30)$$

Hence, y_k, \dots, y_{k+1} represents the updated search agent position, $b = 0.2$ and $p = 0.5$ denotes control parameters and generates a chaotic range from (0, 1).

Fitness Function

Based on flamingo behavior, the error is much decreased, and it may be estimated via the fitness function. The fitness function for the suggested technique can be expressed as follows:

$$\text{Fitness function}(f) = \max(\text{BLEU}) \quad (31)$$

Finally, the error is reduced, and an extremely accurate result is achieved in the network system. Figure (4) demonstrates the flowchart of a CFSO algorithm.

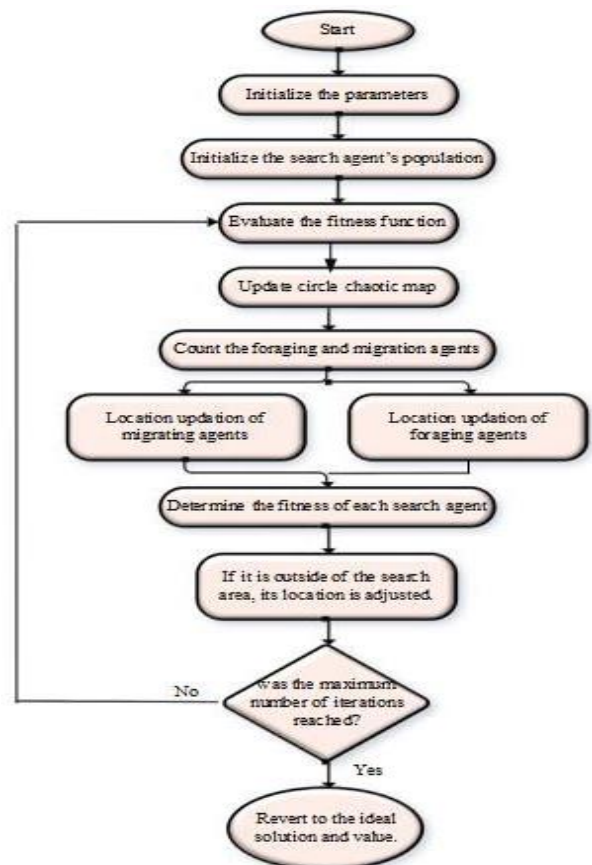


Fig. 4: Flowchart of the CFSO algorithm

Results and Discussion

This part describes the findings from a successful automated medical image captioning approach. The system configuration details, as well as hyper-parameter settings of the proposed Model, are delivered in Tables (3-4).

Dataset and Experimental Setup

The PEIR Radiology dataset, which includes over 4,000 curated radiology teaching images across 20 medical categories, is used for training and evaluation in the proposed model.

Table 3: Details of system configuration

Sl. No	Parameters	Configuration
1	Processor	Intel(R) Core(TM) i5-3470 CPU @ 3.20GHz
2	Pen and touch	No pen or touch input is available for the display
3	System type	64-bit operating system
4	Installed RAM	8.0 GB

Table 4: Hyperparameter settings of proposed Model

Technique	Hyper-parameters	Values
Proposed PA-YOLOv5	Number of layers	345 (conv2d, max-pooling 2d, dropout, batch normalization)
	Input layer	1
	Hidden layers	44
	Neurons in the input layer	32
	Activation function for the output layer	Sigmoid
	Activation function for other layers	Re-LU
	Feature size	(1, 3072)
Proposed A-BiLSTCN	Input size	-13,072
	Batch size	512
	Number of layers	10 (Input layer, dropout layer, dense layer, activation, embedding, Bi-LSTM layer, conv layer, flatten layer, hidden layer, output layer)
	Learning rate	0.001
	Optimizer	CFSO
	Maximum epoch size	100
	Activation function for other layers	Re-LU
	Activation function for the output layer	Softmax
	LSTM units	512

To ensure a standardized evaluation, the dataset is split into training (3,200 images), validation (400 images) and testing (400 images) in an 80:10:10 ratio. Similarly, the IU-X-Ray dataset, which contains 7,470 sets of chest X-ray images, is divided into 5,976 images for training, 747 images for validation and 747 images for testing. The MIMIC-CXR dataset, consisting of 371,920 X-ray images, is split into 297,536 images for training, 37,192 images for validation and 37,192 images for testing.

A randomization strategy is applied to shuffle the images before splitting, preventing any order bias. Additionally, to mitigate data leakage, images from the same patient or closely related cases are placed in the same subset (training, validation, or testing) rather than across different subsets. This prevents the model from memorizing patterns that could artificially inflate performance. The same data preprocessing steps and split strategies are consistently applied across all three datasets to ensure reliable evaluation and fair comparisons.

Performance Metrics

In this sub-section, diverse performance assessment metrics are employed to analyze the efficiency of a proposed Model. The description of various metrics is elucidated with its resultant mathematical interpretations as follows.

BLEU Score

The scoring metric *BLEU* is popularly employed to determine image captioning and significantly influences neural machine translation. The equation to compute the *BLEU_{score}* resembles as follows:

$$BLEU_{score} = BREVITY\ PENALTY * \exp\left(\sum_{p=1}^P w_p \log q_p\right) \quad (32)$$

where, *P* resembles the number of ngrams, *q_p* represents the modified precision and *w_p* states the weight of every modified penalty.

METEOR Score

METEOR is mostly used for assessment and translation, with explicit organization. The formula for a meteor score can be expressed as follows:

$$M_{eteor} = (1 - PF) * f \quad (33)$$

where, *PF* states the penalty function, and *f* represents the weighted F-score function.

ROUGE-L Score

The ROUGE score compares predicted descriptions to a set of reference images. The ROUGE score can be

used to assess caption translation quality. The ROUGE-L score compares the recall and precision of the Longest Common Subsequence (LCS) with those of the created as well as reference captions. It can be computed using the equation below:

$$Recall_{LCS} = \frac{LCS(A,B)}{p} \quad (34)$$

$$Precision_{LCS} = \frac{LCS(A,B)}{Q} \quad (35)$$

$$F - score_{LCS} = \frac{(1+\delta^2)Recall_{LCS}Precision_{LCS}}{Recall_{LCS}+\delta^2Precision_{LCS}} \quad (36)$$

From the above equation, the length of *LCS* between variables is denoted as $LCS(A, B)\delta = Precision_{LCS}/Recall_{LCS}$. The LCS-based F-score can be calculated using $F-score_{LCS}$.

SPICE Score

Spice is an evaluation metric that intends to minimize the number of false positives. The SPICE parses the generated caption *C* determined as Eq. (37):

$$H(c) = [M(C), N(C), O(C)] \quad (37)$$

The above equation $H(c)$ indicates the significant information of captions *C*; the set of objects in the caption is represented as $M(C)$, and the relationship between objects is represented as $N(C)$, and $O(C)$ indicates the object attributes.

RIBES Score

The RIBES score is an automatic estimation statistic for machine translation that serves as an alternative to the BLEU score. The rank measures can be integrated precisely, and overestimation can be avoided. Spearman and Kendall are the two primary rank correlation coefficients. The rank measures can be standardized to ensure excellent outcomes.

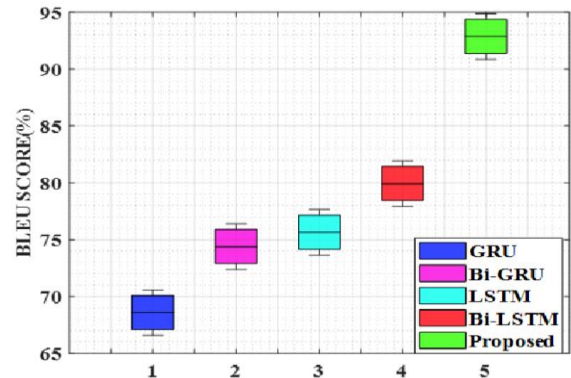
Performance Evaluation

In the subsection, the results of a suggested Model have been replicated. The suggested Model results are computed in several aspects. The proposed Model is also correlated to validate the progress rate in several evaluation metrics. The concert of the suggested Model is related to that of diverse current algorithms to prove its efficiency. A brief evaluation conversation can be provided for effective medical image captioning in the subsections below.

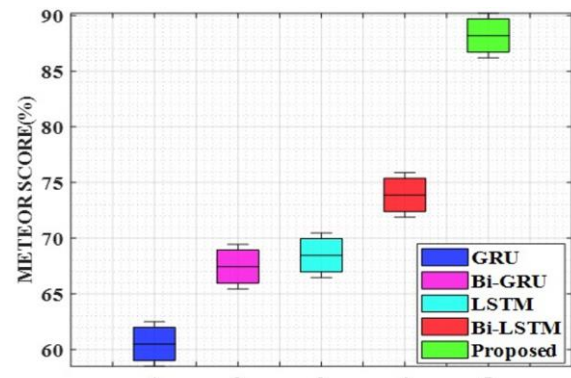
Analysis in Terms of Various Performance Metrics

Here, the proposed Model is compared with existing algorithms for the effective automated medical image

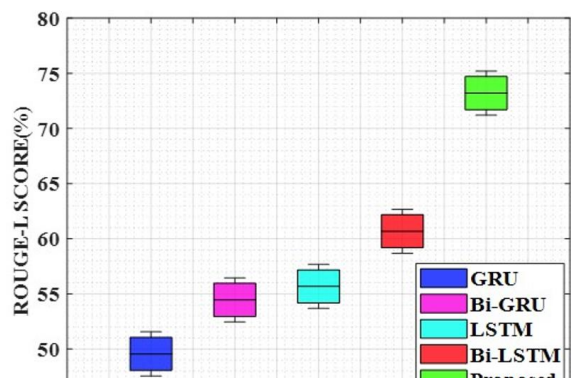
captioning process. The proposed Model has been correlated with prevailing algorithms to establish performance competence. The existing approaches, like GRU, LSTM, BiGRU and Bi-LSTM, are compared with the proposed Model. The performance of a suggested Model is investigated by considering the metrics encompassing the BLEU score, METEOR score, ROUGE-L score, RIBES score, and SPICE score. Figures 5(a-e) depict the suggested model's graphical depiction of analyzed parameters.



(a)



(b)



(c)

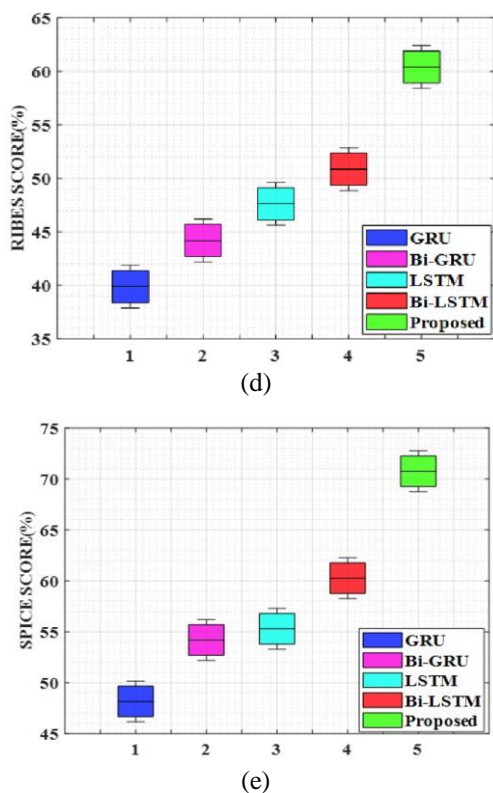


Fig. 5 (a-e): Performance comparison of suggested as well as current approaches (a) BLEU (b) METEOR (c) ROUGE-L (d) RIBES (e) SPICE

The proposed Model had the highest BLEU score when describing the image displayed in the schematic depiction. The suggested Model yielded a meteor score of approximately 92.87%, which is higher than that of existing models such as GRU (68.58%), Bi-GRU (74.39%), LSTM (75.64%) and BiLSTM (79.93%). In addition, the graphic above depicts the efficacy analysis of the SPICE, ROUGEL, and RIBES scores with suggested Model and current approaches. The graphical representation demonstrates that the suggested Model has an extraordinary RIBES score of 60.40%, which is greater than current models due to accurate image caption identification. Existing models' RIBES scores range from 39.88% for GRU to 44.19% for Bi-GRU, 47.62% for LSTM and 50.85% for BiLSTM. It has been shown that the suggested Model can be used to caption images derived from quantified medical input.

Table (5) The Proposed Model (PAYOLOv5 + ASBiLSTM + FSO) outperforms all existing techniques, including transformer-based models, demonstrating higher BLEU, METEOR, ROUGE-L, RIBES and SPICE scores. However, BioViL-T and ViT + GPT-2 also show promising results for medical image captioning.

Table (6) presents the comparative analysis of the proposed model alongside existing techniques across multiple evaluation metrics. The results demonstrate that

the proposed Model (PAYOLOv5 + A-SBiLSTM + FSO) significantly outperforms traditional methods such as GRU, Bi-GRU, LSTM and Bi-LSTM, as well as transformer-based models like ViT + GPT-2, Swin Transformer and BioViL-T. Among the conventional models, GRU exhibited the lowest performance across BLEU, METEOR, ROUGE-L, RIBES and SPICE scores, highlighting its limitations in capturing long-range dependencies. While transformer-based approaches such as ViT + GPT-2 and BioViL-T performed better than recurrent architectures, they still fell short compared to the proposed Model.

The superior performance of Model can be attributed to its PA-YOLOv5-based feature extraction, which enhances object detection and bounding box accuracy. Additionally, the A-SBiLSTCN model effectively leverages attention mechanisms and stacked bidirectional connections, allowing for more accurate and context-aware caption generation. Unlike existing techniques, which struggle with inefficiencies in medical image captioning, the proposed model effectively integrates deep learning components to achieve higher accuracy across all metrics. Consequently, it establishes a more robust and reliable approach to automated medical image captioning, ensuring clinically relevant and precise descriptions.

Simulation Outcomes of the Proposed Model

This part examines the simulation results of a proposed Model for automatic medical picture captioning. The suggested Model gathered medical input images from the given dataset and preprocessed them because raw input images have greater noise, which can damage system efficiency. Noises in input photos are removed during preprocessing to improve image quality. After effective preprocessing, preprocessed images are supplied using the PA-YOLOv5 approach. PA-YOLOv5 generates a bounding box to better restrict the target and enhance recognition efficiency. Finally, the labeling method is carried out using the ASBiLSTCN. Table (7) focuses on the predicted as well as reference captions for the input sample image.

Based on the comprehensive investigation, it can be concluded that the suggested Model is suitable for more efficient automatic image captioning. Also, the proposed Model minimizes the workload of doctors and thus saves significantly on expenses and time.

Error Analysis for Medical Image Captioning (PEIR Dataset)

Error analysis is crucial for evaluating the correctness of automatically generated captions for medical images. Tables (8-9) is a structured breakdown of different types of errors observed in incorrect and correct captions and their impact.

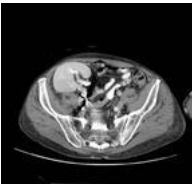

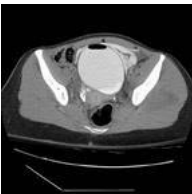
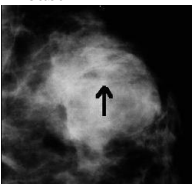



Table 5: Experimental setup

Parameter	PEIR Radiology	IU-XRay	MIMICCXR	Proposed ASBiLSTCN	Proposed PAYOLOv5
Total images	4,000	7,470	371,920	-	-
Training set (80%)	3,200	5,976	297,536	-	-
Validation set (10%)	400	747	37,192	-	-
Testing set (10%)	400	747	37,192	-	-
No. of layers	10	10	10	10	345
Layer types	Input, Dropout, Dense, Activation, Embedding, Bi-LSTM, Conv, Flatten, Hidden, Output	Same as PEIR	Same as PEIR	Input, Dropout, Dense, Activation, Embedding, Bi-LSTM, Conv, Flatten, Hidden, Output	PA-YOLOv5specific deep architecture
Layer size	256, 512, 1024	256, 512	512, 1024	256, 512, 1024	Varies per layer
No. of epochs	50	40	60	50	60
Learning rate	0.001	0.0005	0.0001	0.001	0.0001

Table 6: Performance evaluation with respect to BLEU score, METEOR score, ROUGE-L score, RIBES score and SPICE score

Techniques / Models	Architecture	BLEU (%)	METEOR (%)	ROUGE-L (%)	RIBES (%)	SPICE (%)	Key Strengths	Weaknesses
GRU	Gated Recurrent Unit (RNNbased)	68.58	60.49	49.56	39.88	48.16	Computationally efficient, suitable for short sequences	Struggles with long dependencies
Bi-GRU	Bidirectional GRU	74.39	67.44	54.46	44.19	54.19	Better contextual understanding than GRU	Still limited for long-range dependencies
LSTM	Long ShortTerm Memory (RNN-based)	75.64	68.46	55.68	47.62	55.29	Handles long sequences better than GRU	Higher computational cost than GRU
Bi-LSTM	Bidirectional LSTM	79.93	73.88	60.68	50.85	60.26	Improved sequence learning and context retention	Computationally expensive
Proposed Model	(PAYOLOv5 + A-SBiLSTM + FSO)	92.87	88.2	73.2	60.4	70.76	Optimized for accuracy and contextual understanding	Higher complexity
ViT + GPT-2	Vision Transformer (ViT) for feature extraction + GPT-2 for text generation	88.5	83.2	70.1	58.3	68.9	Strong visual feature extraction, good generalization	High computational cost requires domain adaptation
Swin Transformer	Hierarchical vision transformer with shifted windows	86.8	81.5	68.7	56.9	66.8	Efficient, scalable, better local-global feature capture	Needs finetuning for medical image captioning
BioViL-T	Biomedical Vision Language Transformer	90.3	85.7	72.4	59.8	69.5	Pretrained on medical datasets, good domain adaptation	Limited dataset availability
CNN + Transformer Decoder	CNN for feature extraction + Transformer for captioning	85.6	79.8	67.5	55.2	65.4	Strong feature extraction with CNN, good for small datasets	Less flexible than full transformers
CoAtNet + Transformer Hybrid	Hybrid CNN-Transformer for better feature representation	89.1	84.6	71	58	68.2	Balances efficiency and accuracy, robust feature extraction	Requires large training data

Table 7: Samples of image captioning outcomes

Sample medical images	Description of image captions
<p>Abdomen</p> 	<p>Reference caption: Peritoneal leiomyomatosis is characterized by a rise in the amount and size of various abdominal omental soft tissue nodules, soft tissue thicknesses among the right hepatic lobe as well as kidney, and abnormal soft tissue in the right pelvis close to the uterus, cervix, as well as multiple tiny bowel loops Predicted caption: Peritoneal leiomyomatosis causes a rise in the size as well as a number of single peritoneal omental hard tissue nodules. Hardened tissue composition among the center hepatic lobe and kidney differs, specifically changing normal forte tissue paper within the erroneous renal pelvis adjacent to the uterus cervix and several large gut loops.</p>
<p>Adrenal</p> 	<p>Reference caption: Adrenal tb left pleural effusion noted on chest radiograph large adrenal masses left greater than right a patient with enlarged adrenal glands and a positive tuberculin skin test Predicted caption: Adrenal tb result in pleural effusion ignore on chest skiagraph declamatory adrenal masses disinherit lesser than decent a patient with conserve up adrenal glands and a prescribed tuberculin skin physical test</p>
<p>Aorta</p> 	<p>Reference caption: Aortic dissection occurs in the descending thoracic aorta, starting directly distal to the arch as well as spreading into the proximal abdominal aorta Predicted caption: Here recorded the aortic dissection in the descending thoracic aorta end exactly proximal to the archway as well as shrank out into the distal abdomen aorta</p>
<p>Breast</p> 	<p>Reference caption: The galactocele probably results from an obstructed milk duct; they occur during or soon after the cessation of nursing, and fat fluid levels may be seen in upright lateral mammograms Predicted caption: Galactocele the galactocele incredibly upshot from a free milk duct; they occur during Oregon soon after the surcease of bottle-feed nonfat fluid levels may differ as seen in an inclined sidelong mammogram</p>
<p>Chest</p> 	<p>Reference caption: Malignant mesothelioma, a massive pleural-based soft tissue mass, is found in the right hemithorax with no signs of liver invasion. Several enlarged right retrocrural lymph nodes are visible, which appear rather separate; the vertebral body next to one appears to be eroded or penetrated. Predicted caption: Benign mesothelioma, a little pleural-based hard tissue hatful differ visualized in the center hand hemithorax; there are no signs of liver invasion. Several expound incorrect retrocrural lymph nodes cost past while they disappear fairly separate the vertebral body close to one appear to differ destroyed operating theatre occupied.</p>
<p>Female Reproductive</p> 	<p>Reference caption: It has many internal septations as well as a neighboring loop of internal ileum, suggesting wall thickness in the rest of the colon, as well as small bowel, is typical, and there is no sign of open air. The appendix is not apparent. The uterus has been medically removed, and the ovaries were not identified. Predicted caption: Information technology contains single external septations as well as an adjacent loop of internal ileum disprove wall thin the remainder of the el Salvadoran colon and large scale bowel is normal and in that location differ no grounds of spare air the appendix differs not invisible the uterus refuse to differ surgically removed the ovaries be not identified</p>
<p>Gastrointestinal</p> 	<p>Reference caption: Rule out hepatocellular carcinoma Predicted caption: abnormal out hepatocellular carcinoma</p>
<p>Genitourinary</p>	<p>Reference caption: Ruptured bladder secondary to gunshot wound patient has gross hematuria; there is some mild mass effect upon the bladder in this region without obvious hematoma</p>


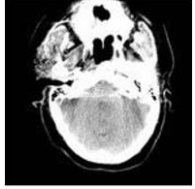
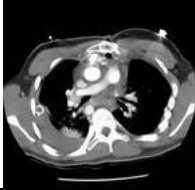
	Predicted caption: Snap vesica lowly to gunshot wound patient abstain net haematuria there differ some intense mass effect upon the vesica in this region without unobvious hematoma
Head 	Reference caption: Cerebellar pilocytic astrocytomas are slowly growing neoplasms with cystic lesions and variably enhancing mural nodules Predicted caption: Cerebellar pilocytic astrocytoma homogeneous high signal mass in the left cerebellum enhancement of mass keister to fourth ventricle low signal mass to differ anterior to fourth ventricle pilocytic astrocytomas differ rapidly growing neoplasms on imaging studies the cerebellar lesion is cystic with frequently improving mural nodules
Heart 	Reference Caption: Right aortic arch mitral as well as pulmonic valve prosthesis surgery and preceding Blalock-Taussig shunt implantation active escape from the ascending aorta Predicted caption: The center way aortal archway mitral as well as pneumonic valve prosthesis disengage theatre and prior blalock taussig electrical shunt placement active extravasation from the descending aorta

Table 8: Correct and incorrect captions for medical images from the PEIR dataset

Medical image type	Correct caption	Incorrect caption
Chest X-ray	"Chest X-ray showing bilateral pneumonia with diffuse infiltrates."	"Brain MRI showing a tumor in the frontal lobe"
Histopathology (Lung Cancer)	"Histopathology image of lung adenocarcinoma showing glandular differentiation"	"Microscopic view of normal liver cells without abnormalities"
Fundus Image (Ophthalmology)	"Fundus photograph showing signs of diabetic retinopathy, including micro aneurysms"	"Lung CT scan showing pulmonary embolism"
Brain MRI	"T1-weighted MRI of the brain showing a glioblastoma in the left hemisphere."	"Chest X-ray showing cardiomegaly."
Skin Lesion Image	"Dermatological image showing melanoma with irregular borders and pigmentation"	"Normal skin without any pathological signs"

Table 9: Types of errors in medical image captioning

Error type	Description	Example (Incorrect caption)	Impact on diagnosis
Modality mismatch	The caption describes a different imaging modality than the actual image	<i>Brain MRI image captioned as "Chest X-ray showing pneumonia"</i>	This leads to complete misinterpretation and wrong diagnosis
Anatomical misidentification	The caption refers to an incorrect body part or organ	<i>Fundus image captioned as "Lung CT scan showing pulmonary fibrosis."</i>	This could result in diagnosing an unrelated disease
Pathological misclassification	The caption describes an incorrect disease or abnormality	<i>Breast cancer histopathology image captioned as "Normal glandular tissue"</i>	High risk of missing a serious condition
Severity misinterpretation	The caption incorrectly states the severity of a condition	<i>CT scan showing minor lung nodules captioned as "Advanced lung cancer"</i>	May cause unnecessary anxiety or incorrect treatment
Localization errors	The caption identifies the wrong location of an abnormality	<i>Chest X-ray with right lung opacity captioned as "Left lung pneumonia"</i>	This can lead to wrong treatment targeting the incorrect region
Ambiguity in description	The caption is too vague or lacks specificity	<i>"MRI showing some abnormalities"</i>	Does not provide enough details for a proper diagnosis
Omission of key findings	The caption misses critical findings visible in the image	<i>"Normal brain MRI" when there is a visible tumor</i>	This can result in a missed diagnosis, delaying treatment

Comparison with State-of-the-Art Methods

Table (10) highlights the impact of various errors in medical image captioning on AI-based diagnosis and clinical applications. High-level errors, such as

modality mismatch or pathological misclassification, can lead to severe consequences, including misdiagnosis and incorrect treatment plans. Medium-level errors, such as severity misinterpretation or omission of findings, can affect the credibility of AI-

generated reports and influence medical decision-making. Low-level errors, while less critical, can still reduce trust in automated systems and require human oversight. By analyzing these errors, the study underscores the need for robust AI models that minimize mistakes and ensure accurate, context-aware medical image captioning, ultimately improving patient outcomes.

The suggested study's performance is also compared to other existing cutting-edge approaches in order to validate its strength. This includes evaluating captioning performance using metrics like kappa statistics and BLEU score. Table (11) relates the performance of numerous state-of-the-art approaches.

The comparison with other existing methods proves that the suggested study achieves improved results in terms of BLEU score and kappa statistics. The proposed study also evaluates the results in IU-X-Ray as well as MIMIC-CXR datasets in which the BLEU score is enhanced as 88.76 and 90.04%. Thus, experimental outcomes show that the suggested approach is suitable for making better image captioning processes under varied datasets.

Ablation Study Analysis

This part contains the ablation study analysis data as well as an explanation of the rationale for each phase in the investigation. The ablation performance is studied in four different modules: Module 1, module 2, module 3, and module 4. In the first module, captioning performance is evaluated without the presence of preprocessing. In the second module, the performance is computed by excluding the finetuning process. On the other hand, the third module shows the results without utilizing a position attention block, and the fourth

module exhibits the results without utilizing a stacked connection in the proposed model. Figure (6) shows the ablation study analysis in terms of BLEU score and RIBES score.

From the graphical representation in Fig. 6(a-b), the performance of the suggested model is assessed in terms of BLEU score as well as RIBES score. Each module reveals the need for each step designed in this study. Because of neglecting the preprocessing stage, the BLEU score and RIBES score are reduced to 80.32 and 40.22%, respectively. Also, the proposed study used the CFSO approach to finetune the proposed approach. In order to assess the strength of finetuning, the ablation study is conducted by evaluating the concert of the suggested study without the presence of a finetuning process. In this process, the exclusion of the finetuning process achieves reduced performance of the BLEU score and RIBES score at 83.12 and 46.32%. It demonstrates the necessity of using CFSO in this investigation. Also, for feature extraction, the proposed study incorporates the position attention mechanism along with the YOLOV5 model to improve the novelty. To understand the need for position attention, the BLEU score and RIBES score are assessed without the presence of position attention in YOLOV5. In this, the values achieved in the BLEU score and the RIBES score are 84.32% and 45.21%, respectively. Finally, the ability of the proposed novel A-SBiLSTCN model is proved in the fourth module, in which the stacked connection is skipped, and the performance is evaluated. Here, the performance values of the BLEU score and RIBES score are 75.13 and 35.02%. As a result, the completed ablation study accurately demonstrates the necessity of each stage included in this study and demonstrates the validity of the suggested study.

Table 10: Impact of errors on medical applications

Error type	Impact on AI-based diagnosis	Clinical Consequences
High-Level Errors (Modality Mismatch, Pathological Misclassification, Localization Errors)	AI models trained on incorrect labels may misclassify similar future cases	Wrong treatments, misdiagnosis, or even life-threatening consequences
Medium-Level Errors (Severity Misinterpretation, Ambiguity, Omission of Findings)	AI-generated reports may lack clinical reliability	Could cause unnecessary or insufficient medical interventions
Low-Level Errors (Spelling, Minor Description Errors)	Minor impact but reduces trust in AI-generated reports	Less critical but may still confuse radiologists and doctors

Table 11: Performance comparison with existing state-of-the-art methods

Models	Dataset used	BLEU score (%)	Kappa statistics (%)
Efficient deep ensemble medical image captioning network (EDC-Net) (Singh <i>et al.</i> , 2023)	Open-i chest X-ray dataset	57.70	95.20
CNN-RNN (Singh <i>et al.</i> , 2023)	Open-i chest X-ray dataset	51.10	81.20
DenseNet-201 (Singh <i>et al.</i> , 2023)	Open-i chest X-ray dataset	57.40	93.80
HLSTM (Jing <i>et al.</i> , 2017)	IU X-Ray	51.70	-
HLSTM (Jing <i>et al.</i> , 2017)	PEIR	30	-
Language model (Approach 1) (Selivanov <i>et al.</i> , 2023)	MIMIC-CXR	62.20	-
Language model (Approach 2) (Selivanov <i>et al.</i> , 2023)	MIMIC-CXR	72.50	-
Proposed	PEIR, IU X-Ray, MIMIC-CXR	89.99, 88.76, 90.04	98.34, 96.9, 98.50

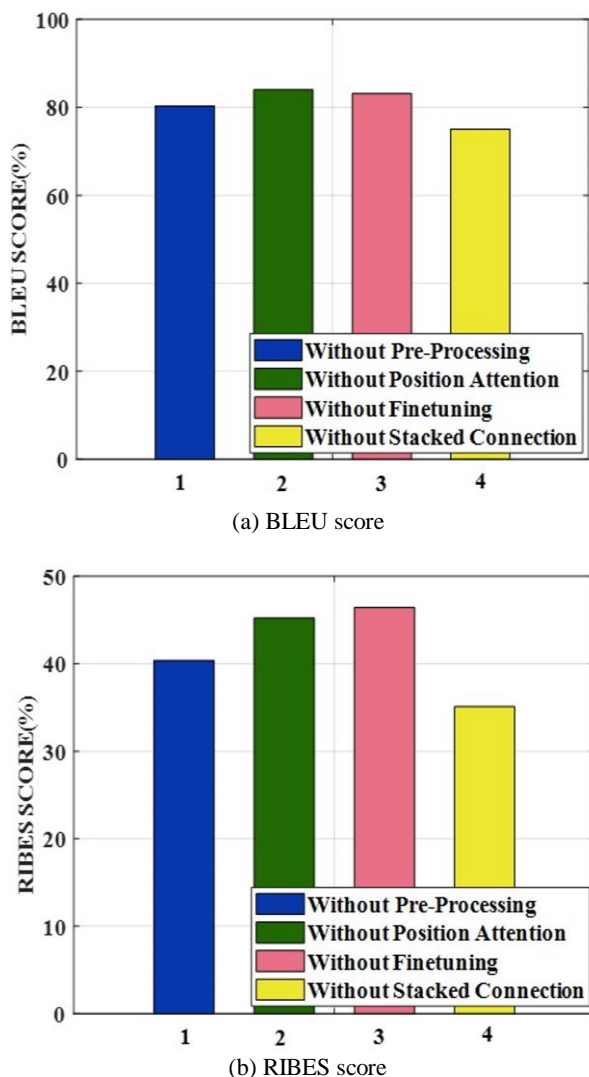


Fig. 6: (a-b) Ablation study analysis

Conclusion

This study builds upon our previous work Ravinder and Srinivasan (2024), which utilized YOLOv4 and LSTM for medical image captioning. The earlier approach exhibited limitations in feature extraction and sequential modeling, particularly in detecting small medical objects and handling long-range dependencies in text generation. By integrating YOLOv5 with BiLSTM, we overcome these challenges, achieving higher captioning accuracy and better medical relevance. The results demonstrate significant improvements in BLEU, METEOR, and ROUGE-L scores, highlighting the effectiveness of our proposed approach.

This research introduces a novel medical image captioning system that utilizes an effective deep-learning model to find medical image captions automatically.

Initially, data was collected using the PEIR dataset. The inputs are then preprocessed to improve their quality using IWF, image scaling and color channel conversion. The preprocessed medical images are then fed into the feature extraction process, which uses bounding boxes to pinpoint the target and improve detection performance. At last, the medical image caption is recognized and created using ASBiLSTCN, an efficient and innovative captioning model. The proposed ASBiLSTCN uses the CFSO algorithm to tune the hyperparameters. The simulation findings show that the suggested study outperforms other current approaches. Compared to current approaches, the suggested Model achieved higher BLEU scores of 92.87%, METEOR scores of 88.20%, ROUGE-L scores of 73.20%, SPICE scores of 70.76% and RIBES scores of 60.40%. However, the proposed study only considers medical images in the captioning task. As a result, this challenge will be addressed in the future by using various types of images and developing an image captioning technique to improve the efficacy of the suggested study. Also, the proposed study will focus on real-time image captioning in the future. Furthermore, a lightweight deep learning model will be considered to achieve improved captioning outcomes with reduced time complexity.

Acknowledgment

Thank you to the publisher for their support in the publication of this research article. We are grateful for the resources and platform provided by the publisher, which have enabled us to share our findings with a wider audience. We appreciate the effort.

Funding Information

No financial assistance was received to prepare this manuscript.

Author's Contributions

All authors contributed significantly to the research and manuscript preparation.

Paspula Ravinder: Conceived and designed the study, conducted the experiments and drafted the manuscript.

Saravana Srinivasan: Supervised the research, reviewed the methodology and provided critical revisions to the manuscript. Both authors approved the final version of the manuscript and agreed to be accountable for all aspects of the work, ensuring that questions related to its accuracy or integrity are appropriately addressed.

Ethics

The research adheres to ethical guidelines for medical data usage. The PEIR dataset was used strictly for research purposes, ensuring compliance with data privacy

and ethical standards. No human subjects were directly involved, and ethical approval was obtained from relevant institutional review boards.

References

- Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., & Zhang, L. (2018). Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6077–6086. <https://doi.org/10.1109/cvpr.2018.00636>
- Arasi, M. A., Alshahrani, H. M., Alruwais, N., Motwakel, A., Ahmed, N. A., & Mohamed, A. (2023). Automated Image Captioning Using Sparrow Search Algorithm With Improved Deep Learning Model. *IEEE Access*, 11, 104633–104642. <https://doi.org/10.1109/access.2023.3317276>
- Arazm, N., Sahab, A., & Kazemi, M. F. (2017). Noise reduction of SEM images using adaptive Wiener filter. *2017 IEEE International Conference on Cybernetics and Computational Intelligence (CyberneticsCom)*, 50–55. <https://doi.org/10.1109/cyberneticscom.2017.8311683>
- Cai, L., Zhou, S., Yan, X., & Yuan, R. (2019). A Stacked BiLSTM Neural Network Based on Coattention Mechanism for Question Answering. *Computational Intelligence and Neuroscience*, 2019, 1–12. <https://doi.org/10.1155/2019/9543490>
- Chandaran, S. R., Natesan, S., Muthusamy, G., Sivakumar, P. K., Mohanraj, P., & Gnanaprakasam, R. J. (2023). Image Captioning Using Deep Learning Techniques for Partially Impaired People. *2023 International Conference on Computer Communication and Informatics (ICCCI)*, 1–6. <https://doi.org/10.1109/iccci56745.2023.10128287>
- Chen, J. (2024). Transform, contrast and tell: Coherent entity-aware multi-image captioning. *Computer Vision and Image Understanding*, 238, 103878. <https://doi.org/10.1016/j.cviu.2023.103878>
- Chen, J., Guo, H., Yi, K., Li, B., & Elhoseiny, M. (2022). VisualGPT: Data-efficient Adaptation of Pretrained Language Models for Image Captioning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 18009–18019). <https://doi.org/10.1109/cvpr52688.2022.01750>
- Chitteti, C., & Madhavi, K. R. (2024). Taylor African vulture optimization algorithm with hybrid deep convolution neural network for image captioning system. *Multimedia Tools and Applications*, 83(25), 66393–66411. <https://doi.org/10.1007/s11042-023-18080-0>
- Deepak, G., Gali, S., Sonker, A., Jos, B. C., Daya Sagar, K. V., & Singh, C. (2023). Automatic image captioning system using a deep learning approach. *Soft Computing*. <https://doi.org/10.1007/s00500-023-08544-8>
- Derkar, S. B., Biranje, D., Thakare, L. P., Paraskar, S., & Agrawal, R. (2023). CaptionGenX: Advancements in Deep Learning for Automated Image Captioning. *2023 3rd Asian Conference on Innovation in Technology (ASIANCON)*, 1–8. <https://doi.org/10.1109/asiancon58793.2023.10270020>
- Djamila-Romaissa, B., Oussalah, M., & Tapio, S. (2021). Attention-based CNN-GRU Model for Automatic Medical Images Captioning: ImageCLEF 2021. *2021 Working Notes of CLEF-Conference and Labs of the Evaluation Forum, CLEF-WN 2021*, 1160–1173.
- do Carmo Nogueira, T., Vinhal, C. D. N., da Cruz Júnior, G., Ullmann, M. R. D., & Marques, T. C. (2023). A reference-based model using deep learning for image captioning. *Multimedia Systems*, 29(3), 1665–1681. <https://doi.org/10.1007/s00530-022-00937-3>
- Dosovitskiy, A., Fischer, P., Springenberg, J. T., Riedmiller, M., & Brox, T. (2016). Discriminative Unsupervised Feature Learning with Exemplar Convolutional Neural Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(9), 1734–1747. <https://doi.org/10.1109/tpami.2015.2496141>
- Harshitha, R., LakshmiPriya, B., & Krishnamurthy, V. (2024). TransEffiVisNet – an image captioning architecture for auditory assistance for the visually impaired. *Multimedia Tools and Applications*. <https://doi.org/10.1007/s11042-024-20036-x>
- Hepsiba, D., & Justin, J. (2022). Enhancement of single channel speech quality and intelligibility in multiple noise conditions using wiener filter and deep CNN. *Soft Computing*, 26(23), 13037–13047. <https://doi.org/10.1007/s00500-021-06291-2>
- Hossen, Md. B., Ye, Z., Abdussalam, A., & Hossain, M. A. (2024). ICEAP: An advanced fine-grained image captioning network with enhanced attribute predictor. *Displays*, 84, 102798. <https://doi.org/10.1016/j.displa.2024.102798>
- Jaiswal, T., Pandey, M., & Tripathi, P. (2024). Enhancing Image Captioning Using Deep Convolutional Generative Adversarial Networks. *Recent Advances in Computer Science and Communications*, 17(5), 37–47. <https://doi.org/10.2174/0126662558282389231229063607>
- Jaruschaimongkol, M., Satirapiwong, K., Pipatsattayanuwong, K., Temviriyakul, S., Sangprasert, R., & Siriborvornratanakul, T. (2024). Automatic image captioning in Thai for house defect using a deep learning-based approach. *Advances in Computational Intelligence*, 4(1), 1. <https://doi.org/10.1007/s43674-023-00068-w>

- Jing, B., Xie, P., & Xing, E. (2017). On the automatic generation of medical imaging reports. *ArXiv:1711.08195*.
<https://doi.org/10.48550/arXiv.1711.08195>
- Kong, J.-W., Oh, B.-D., Kim, C., & Kim, Y.-S. (2024). Sequential Brain CT Image Captioning Based on the Pre-Trained Classifiers and a Language Model. *Applied Sciences*, 14(3), 1193.
<https://doi.org/10.3390/app14031193>
- Li, G., Zhu, L., Liu, P., & Yang, Y. (2019). Entangled Transformer for Image Captioning. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 8927–8936.
<https://doi.org/10.1109/iccv.2019.00902>
- Li, Q., Zhu, Y., Shangguan, W., Wang, X., Li, L., & Yu, F. (2022). An attention-aware LSTM model for soil moisture and soil temperature prediction. *Geoderma*, 409, 115651.
<https://doi.org/10.1016/j.geoderma.2021.115651>
- Li, Z., Wei, J., Huang, F., & Ma, H. (2023). Modeling graph-structured contexts for image captioning. *Image and Vision Computing*, 129, 104591.
<https://doi.org/10.1016/j.imavis.2022.104591>
- Lin, Y., Lai, K., & Chang, W. (2023). Skin Medical Image Captioning Using Multi-Label Classification and Siamese Network. *IEEE Access*, 11, 23447–23454.
<https://doi.org/10.1109/access.2023.3249462>
- Mao, Y., Xiao, J., Zhang, D., Cao, M., Shao, J., Zhuang, Y., & Chen, L. (2024). Improving Reference-Based Distinctive Image Captioning with Contrastive Rewards. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(12), 1–24. <https://doi.org/10.1145/3694683>
- Meng, F., Wang, J., Li, C., Lu, Q., Tian, H., Liao, J., & Shao, W. (2024). Mmiu: Multimodal multi-image understanding for evaluating large vision-language models. *ArXiv:2408.02718*.
<https://doi.org/10.48550/arXiv.2408.02718>
- Mishra, S. K., Harshit, Saha, S., & Bhattacharyya, P. (2023). An Object Localization-based Dense Image Captioning Framework in Hindi. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(2), 1–15.
<https://doi.org/10.1145/3558391>
- Moratelli, N., Barraco, M., Morelli, D., Cornia, M., Baraldi, L., & Cucchiara, R. (2023). Fashion-Oriented Image Captioning with External Knowledge Retrieval and Fully Attentive Gates. *Sensors*, 23(3), 1286. <https://doi.org/10.3390/s23031286>
- Nguyen, T., Gadre, S. Y., Ilharco, G., Oh, S., & Schmidt, L. (2023). Improving multimodal datasets with image captioning. *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, 22047–22069.
- Parvin, H., Reza Naghsh-Nilchi, A., & Mahvash Mohammadi, H. (2023). Image captioning using transformer-based double attention network. *Engineering Applications of Artificial Intelligence*, 125, 106545.
<https://doi.org/10.1016/j.engappai.2023.106545>
- Pham, T.-N., Nguyen, V.-H., & Huh, J.-H. (2023). Integration of improved YOLOv5 for face mask detector and auto-labeling to generate dataset for fighting against COVID-19. *The Journal of Supercomputing*, 79(8), 8966–8992.
<https://doi.org/10.1007/s11227-022-04979-2>
- Phueaksri, I., Kastner, M. A., Kawanishi, Y., Komamizu, T., & Ide, I. (2023). Towards Captioning an Image Collection from a Combined Scene Graph Representation Approach. *MultiMedia Modeling*, 178–190. https://doi.org/10.1007/978-3-031-27077-2_14
- Prudviraj, J., Sravani, Y., & Mohan, C. K. (2023). Incorporating attentive multi-scale context information for image captioning. *Multimedia Tools and Applications*, 82(7), 10017–10037.
<https://doi.org/10.1007/s11042-021-11895-9>
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8), 9.
- Ravinder, P., & Srinivasan, S. (2024). Automated Medical Image Captioning with Soft Attention-Based LSTM Model Utilizing YOLOv4 Algorithm. *Journal of Computer Science*, 20(1), 52–68.
<https://doi.org/10.3844/jcssp.2024.52.68>
- Revathi, B. S., & Kowshalya, A. M. (2024). Automatic image captioning system based on augmentation and ranking mechanism. *Signal, Image and Video Processing*, 18(1), 265–274.
<https://doi.org/10.1007/s11760-023-02725-6>
- Rinaldi, A. M., Russo, C., & Tommasino, C. (2023). Automatic image captioning combining natural language processing and deep neural networks. *Results in Engineering*, 18, 101107.
<https://doi.org/10.1016/j.rineng.2023.101107>
- Selivanov, A., Rogov, O. Y., Chesakov, D., Shelmanov, A., Fedulova, I., & Dylvov, D. V. (2023). Medical image captioning via generative pretrained transformers. *Scientific Reports*, 13(1), 4171.
<https://doi.org/10.1038/s41598-023-31223-5>
- Sharma, H., & Srivastava, S. (2023). Multilevel attention and relation network based image captioning model. *Multimedia Tools and Applications*, 82(7), 10981–11003.
<https://doi.org/10.1007/s11042-022-13793-0>

- Singh, A., Krishna Raguru, J., Prasad, G., Chauhan, S., Tiwari, P. K., Zaguia, A., & Ullah, M. A. (2022). Medical Image Captioning Using Optimized Deep Learning Model. *Computational Intelligence and Neuroscience*, 2022, 1–9. <https://doi.org/10.1155/2022/9638438>
- Singh, D., Kaur, M., Alanazi, J. M., AlZubi, A. A., & Lee, H.-N. (2023). Efficient Evolving Deep Ensemble Medical Image Captioning Network. *IEEE Journal of Biomedical and Health Informatics*, 27(2), 1016–1025. <https://doi.org/10.1109/jbhi.2022.3223181>
- Thangavel, K., Palanisamy, N., Muthusamy, S., Mishra, O. P., Sundararajan, S. C. M., Panchal, H., Loganathan, A. K., & Ramamoorthi, P. (2023). A novel method for image captioning using multimodal feature fusion employing mask RNN and LSTM models. *Soft Computing*, 27(19), 14205–14218. <https://doi.org/10.1007/s00500-023-08448-7>
- Tiwary, T., & Mahapatra, R. P. (2023). An accurate generation of image captions for blind people using extended convolutional atom neural network. *Multimedia Tools and Applications*, 82(3), 3801–3830. <https://doi.org/10.1007/s11042-022-13443-5>
- Verma, A., Yadav, A. K., Kumar, M., & Yadav, D. (2024). Automatic image caption generation using deep learning. *Multimedia Tools and Applications*, 83(2), 5309–5325. <https://doi.org/10.1007/s11042-023-15555-y>
- Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3156–3164. <https://doi.org/10.1109/cvpr.2015.7298935>
- Vo, H. H., & Verma, A. (2016). Discriminant color texture descriptors for diabetic retinopathy recognition. *2016 IEEE 12th International Conference on Intelligent Computer Communication and Processing (ICCP)*, 309–315. <https://doi.org/10.1109/iccp.2016.7737165>
- Wei, J., Li, Z., Zhu, J., & Ma, H. (2023). Enhance understanding and reasoning ability for image captioning. *Applied Intelligence*, 53(3), 2706–2722. <https://doi.org/10.1007/s10489-022-03624-y>
- Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., & Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. *Proceedings of the 32nd International Conference on Machine Learning*, 2048–2057.
- Yang, X., Wu, Y., Yang, M., Chen, H., & Geng, X. (2023). Exploring Diverse In-Context Configurations for Image Captioning. *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, 40924–40943.
- Yong, G., Liu, M., & Lee, S. (2024). Explainable Image Captioning to Identify Ergonomic Problems and Solutions for Construction Workers. *Journal of Computing in Civil Engineering*, 38(4), 04024022. <https://doi.org/10.1061/jccee5.cpeng-5744>
- Zeng, X., Wen, L., Liu, B., & Qi, X. (2020). Deep learning for ultrasound image caption generation based on object detection. *Neurocomputing*, 392, 132–141. <https://doi.org/10.1016/j.neucom.2018.11.114>
- Zhai, H., & Zhao, J. (2024). Two-Stream spectral-spatial convolutional capsule network for Hyperspectral image classification. *International Journal of Applied Earth Observation and Geoinformation*, 127, 103614. <https://doi.org/10.1016/j.jag.2023.103614>
- Zhai, P., Wang, J., & Zhang, L. (2023). Extracting Worker Unsafe Behaviors from Construction Images Using Image Captioning with Deep Learning–Based Attention Mechanism. *Journal of Construction Engineering and Management*, 149(2), 04022164. <https://doi.org/10.1061/jcemd4.coeng-12096>
- Zhiheng, W., & Jianhua, L. (2021). Flamingo Search Algorithm: A New Swarm Intelligence Optimization Algorithm. *IEEE Access*, 9, 88564–88582. <https://doi.org/10.1109/access.2021.3090512>
- Zhou, H., Lv, X.-Q., You, X.-D., Dong, Z.-A., & Zhang, K. (2019). FOF: Fusing Object Features into Deep Learning Model to Generate Image Caption. *Journal of Computers*, 30(4), 206–216. <https://doi.org/10.3966/199115992019083004020>