MASNET-ESN: An Effective Deep Learning Based Automatic Medical Image Captioning

V S RATNA KUMARI A
School of Computer Science and Engineering
VIT-AP University, India
ratna.23PHD7109@vitap.ac.in

Dr. Lalitha Kumari Pappala
School of Computer Science and Engineering
VIT-AP University, India
lalitha.p@vitap.ac.in

Abstract: *These days, one of the most well-known fields is medical picture captioning. Medical image interpretation and captioning can be expensive and time-consuming, and they frequently call for professional assistance. It is becoming more difficult for radiologists to manage their tasks independently due to the increasing volume of medical images. Therefore, automating the medical image captioning process is introduced to alleviate the high cost and time difficulties while supporting radiologists in enhancing the dependability and precision of the generated captions. Additionally, it offers less experienced new radiologists the chance to take advantage of automatic support. However, the previous studies contain several unsolved challenges, such as producing excessively elaborate captions, having trouble identifying aberrant regions in complicated images, and having low accuracy. To address these issues, we suggest an effective Multiscale Attention Siamese Network (MASNet)-Echo State Network (ESN) based deep learning techniques for automatic medical image captioning. In this work, MASNet extracts global and local visual characteristics from the preprocessed image. Afterwards, ESN can use the image's retrieved high and low-level characteristics to produce a complete description of the input image. Moreover, Tunicate Swarm Algorithm (TSA) based hyperparameter tuning is applied to improve the performance of the ESN network. Finally, the suggested technique will be measured by metrics like Consensus-based Image Description Evaluation (CIDEr), Metric for Evaluation of Translation with Explicit Ordering (METEOR), Recall-Oriented Understudy for Gisting Evaluation-Longest Common Subsequence (ROUGE-L), and Bilingual Evaluation Understudy (BLEU) on the Pathology Education Instructional Resource-Gross (PEIR-Gross), Medical Information Mart for Intensive Care-Chest X-Ray (MIMIC CXR), Radiology Objects in Context (ROCO), and Indiana University chest X-ray (IU X-ray) datasets, and the findings will be compared with other previous methods*.

Keywords: *Medical image, MASNet, ESN, deep learning, TSA*.

Received February 5, 2025; accepted September 21, 2025
https://doi.org/10.34028/iajit/23/2/10

1. Introduction

In today's diagnostic workflows, medical imaging is essential. Medical practitioners and radiologists utilize medical pictures for disease diagnosis and treatment, among other widespread applications in the medical field. Pharmacists may utilize them to find new drugs, and surgeons could use imaging to track the course of treatment before, after, and during a surgical procedure [7, 29, 44]. After analyzing these medical pictures, skilled medical professionals manually create text reports that include their abnormal, regular, or potentially abnormal findings in detailed paragraphs. This report explains the characteristics of the probable condition and provides an overall image summary [8, 26, 47]. Furthermore, the reports exhibit comparable trends for pneumothorax, pleural effusion, heart, pulmonary, and bone structural abnormalities.

For analysts without prior expertise or training, writing medical documentation in text format can be tricky since it necessitates a close inspection of the photographs and a thorough comprehension of the medical situation [9, 14, 27]. Furthermore, this work is tedious and time-consuming for seasoned medical professionals because they must review numerous medical images every day, and it takes them at least thirty minutes to analyze an image and document their findings. Therefore, it is a tedious and uncomfortable task for medical experts. Furthermore, the usual manual annotation overload can result in several issues, including missed discoveries, uneven quantification, and extended hospital stays for patients, all of which raise the expense of care. In addition, labor pressure grows dramatically due to the lack of medical experts in a country with a large population like Pakistan [19, 31, 45]. Additionally, the percentage of incorrect diagnoses is higher in areas lacking in medical treatment facilities. It is problematic, especially in low-income countries where patients might have to pay more to see doctors and obtain medications again.

To solve these issues, an automatic image captioning system is developed to help with the medical image reporting process. These systems automatically obtain details from medical pictures and provide a written report with detailed information, much like a doctor [4, 6, 50]. It lessens the physician's time to manually analyze features from photos and create a written report, which lightens their workload. The computerized and efficient generation of medical reports also reduces the

number of professionals required to compile reports.

Several approaches have been proposed earlier to tackle the task of automatic medical image captioning. However, they are affected by inconsistency and poor accuracy [5, 30, 39]. Two factors cause this: first, the complicated anatomic frameworks that are present in the medical images; second, rather than mapping every object in the picture as in a standard image captioning system; the descriptions ought to concentrate only on aspects that are clinically essential and pertinent to the diagnosis [38, 48]. In addition, current methods view picture description as a standalone process and disregard how context affects word choice. Consequently, they cannot preserve contextual consistency among the lengthy paragraphs that make up medical reports.

Therefore, we introduced a deep-leaning framework for automatic medical image captioning in this work. In this framework, the Multiscale Attention Siamese Network (MASNet) utilized the preprocessed medical image as the input and performs the feature extraction process. Using pooling and multilayer convolution processes, it independently learns and extracts all local and global features from the data, producing an abstract feature mapping that is more accurate than conventional feature extraction techniques. Then, Echo State Network (ESN) examined the retrieved features and produced the image's caption. ESNs use reservoir computing, which generates complex, nonlinear dynamics from a vast pool of recurrently coupled neurons. It allows recording complex patterns and long-range dependencies in sequential data, such as captions.

The important contributions of this paper are as follows:

- We present MASNet, an effective visual extractor that integrates the attention mechanisms that allow the network to ignore noisy or irrelevant portions of the input data and concentrate on pertinent portions.
- An ESN-based caption-generating model is introduced to generate precise results by effectively learning the characteristics and producing appropriate medical captioning.
- To reduce network overfitting and improve the results, Tunicate Swarm Algorithm (TSA) based hyperparameter tuning is implemented.
- Several relevant tests are conducted on the Medical Information Mart for Intensive Care-Chest X-Ray (MIMIC CXR), Indiana University Chest X-ray (IU-X-ray), Radiology Objects in Context (ROCO) and Pathology Education Instructional Resource-Gross (PEIR-Gross) datasets, and the suggested strategy outperformed the earlier models according to data analysis of different performance measures.

This paper's overview is organized as follows: section 1 introduces the work and emphasizes its contributions and motivation. The latest related works on creating image descriptions are discussed in section 2. Section 3 provides a thorough description of the suggested methodology. A thorough explanation of implementation and the outcomes is given in Section 4. The work provided is concluded in section 5.

2. Literature Review

A unique method for captioning skin medical images was presented by Lin *et al*. [15]. In this work, a three-stage framework was implemented to handle the whole process of image captioning. Initially, the characteristics from the auto encoder and discriminator were retrieved using comparable fully convolutional network topologies once the training procedures were finished. Subsequently, the multi-label classifier determined the correlation between the input image and significant keywords that comprised the crucial details about the image. Finally, the Siamese network established a relationship among the sentence specifications and the classifier's keywords.

In another work, Kim *et al*. [11] proposed a distil GPT-2 and Three-Dimensional Convolutional Neural Network (3D-CNN) model for Computed Tomography (CT) image captioning. In this, 3D-CNN was utilized as the encoder to analyze the input image and distilGPT-2 was utilized as the decoder to produce the image captions. Finally, the authors compared and evaluated the results of four combinations of language models and 3D-CNN models. Furthermore, the effects of training with different penalty values and adding penalties to the loss function were investigated. Finally, this picture captioning model attained a 0.35 BLEU value.

Similarly, Kong *et al*. [12] introduced a technique to produce text for consecutively recorded Intracerebral Haemorrhage (ICH) CT images using GPT-2 and CNN classifier. First, CNN was fine-tuned to detect ICH in single CT images that are publically available. Then, it extracted the feature vectors, or matrices, from ICH CT pictures. Afterwards, GPT-2 receives these vectors and text inputs and is trained to produce descriptions for a series of CT scans. Finally, during the experimental evaluation, the authors assessed the four models' performance and identified that the DenseNet121 and ReseNet50V2 models achieved high scores in the N-gram-based technique.

An automatic picture captioning model was proposed by Tiwary and Mahapatra [36]. It consists of the following steps: gathering data, choosing non-captioned images, retrieval of visual and textured characteristics, and the creation of automated captions for the images. First, information was gathered from two open sources, and Adaptive Rain Optimization was used to select photos without captions. Following that, the texture features were retrieved using a weighted patch local binary pattern, and the appearance characteristic was extracted using the spatial derivative and multiscale technique. Lastly, an Extended Convolutional Atom Neural Network (ECANN) automatically created the

captions. To conduct the caption reusing system and choose the best correct caption, the ECANN combined the Long Short-Term Memory (LSTM) and CNN techniques.

Tan *et al*. [34] introduced the multimodal data-assisted knowledge fusion network for a medical image description that uses multimodal auxiliary signals to direct the transformer network towards accurate captions. Audio auxiliary signals specifically offer distinct aberrant visual regions to address the issue of visual data bias. Nevertheless, similar-sounding audio modality signals are not recognizable, which could result in inaccurate audio label mapping to medical imaging regions. As a result, to enhance the model's overall performance, they merge textual aspects with audio. Ultimately, the study involved conducting experiments on two medical image description datasets, namely COVID-19 CT Report (COV-CTR) and IU-X-ray, and analyzing the effectiveness of the suggested methodology.

Selivanov *et al*. [28] presented two structures for captioning X-ray images. This approach used the GPT-3 language model to increase the accuracy of clinical records produced by the encoder-decoder. An extensive medical report is written by the GPT-3 using the information provided by the Encoder with LSTM to identify diseases and highlight areas requiring more care. The model was evaluated using natural language assessment metrics on three medical datasets: MIMIC-CXR, Open-I, and MS-COCO.

Ayesha *et al*. [2] used the deep learning-based deep encoder-decoder architecture to create captions for medical imagery. This framework used a pre-trained deep features extractor named ChexNet to extract the top ten features from the input data. Afterwards, these top ten features were forwarded to the recurrent neural network for caption generation. To ensure its ability, the authors conducted numerous trials on the chest X-ray datasets, and it achieved 0.307 for the BLEU-1 metric.

3. Proposed Methodology

Figure 1 presents an overview of the suggested model for developing medical image captioning. The framework has three components: 1) Preprocessing images and captions, 2) MASNet-based feature extraction, and 3) ESN-based caption generation. Initially, preprocessing methods are used to prepare the input image and captions for further analysis. Then, high and low-level visual characteristics are extracted from the input image using a feature extractor based on MASNet. MASNet's multiscale attention mechanisms concentrate on various granularities within the incoming data. It improves the contextual comprehension of the model by simultaneously considering unique features and more general contextual information. Then, ESN takes the extracted multimodal features and preprocessed captions as the input and generates the relevant medical caption. Its reservoir dynamics and recurrent connections allow it to give an organized representation of the input captions. Within the text data, this organized form captures relationships and temporal information. Additionally, TSA-based hyperparameter adjustment is used to enhance the ESN's performance.

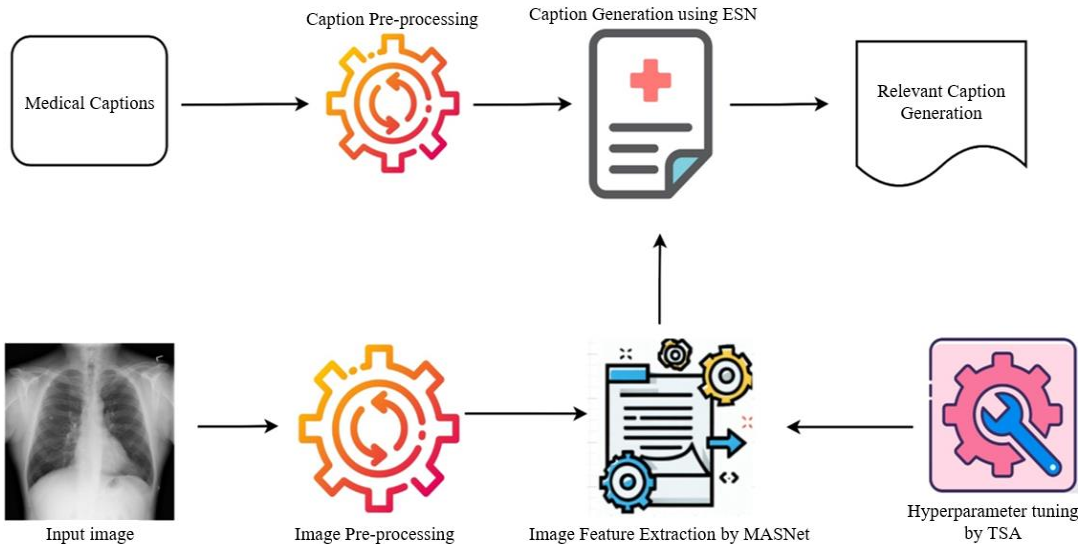


Figure 1. System architecture.

3.1. Preprocessing Stage

When using machine or deep learning techniques, the accuracy of the model predictions must be increased through data preparation. As part of our work, we must prepare the images and captions using the following preprocessing steps for further processing.

3.1.1. Image Preprocessing

This study's datasets, PEIR-Gross, IU X-ray, ROCO, and MIMIC-CXR, come from a variety of sources, including clinical radiographs and educational pathology images. Additionally, the datasets differ in modality, varying from grayscale chest X-rays and

multi-modal radiology pictures (CT, MRI, and ultrasound) in MIMIC-CXR and ROCO to photographic pathology images in PEIR-Gross. As a result, there are visual artifacts in all datasets, including arrows, text labels, superimposed annotations, and inconsistent resolution. Therefore, to ensure consistency, the images were resized to a standard pixel size of 256*256. The pictures in the PEIR-Gross collection were then transformed into grayscale versions because the IU XRay, ROCO, and MIMIC CXR datasets contains greyscale images. Then, during model training, picture normalization was applied to get the best results. Since the intensity range of the input images is more extensive (0-255), deep model weights are frequently initialized with tiny random values (0-1). To reduce issues like bursting gradients that might hinder learning, the input images were normalized within the 0-1 range. Contrast Limited Adaptive Histogram Equalization (CLAHE) was used to further increase local contrast and visual clarity, especially in areas with poor dynamic range. These preprocessing procedures were designed to reduce the inconsistencies caused by artifacts and acquisition variability, guaranteeing consistency throughout model training. Moreover, image augmentation techniques, including rotation, vertical flipping, and horizontal flipping, are performed to boost data samples and decrease network overfitting.

3.1.2. Caption Preprocessing

All numbers (0-9), unique characters, and punctuation were removed from picture captions. Then, all captions have been changed to lowercase to maintain uniformity. The model was assisted in identifying sentence boundaries by introducing unique tokens, "startseq" and "endseq," at the beginning and end of captions. Tokenization was used to divide the report into words and give each word a distinct number token. Each distinct word was allocated a token number and stored as an array named vocabulary. Word embedding was completed outside the model by assigning a unique token number to each word and creating a word embedding matrix that the deep learning methods could utilize because it requires numerical input.

3.2. MASNet-Based Feature Extractor

After preprocessing, the input picture is sent to the MASNet for visual and semantic feature extraction. The network can concentrate on various image regions and gather pertinent visual and semantic data with the help of the multiscale attention mechanism. This network contains three main parts. They are a Siamese CNN feature extractor, a prediction unit and a Multi-Scale Attention (MSA) unit. The Siamese ResNet-50 is the backbone architecture of this network, omitting the global pooling layer and fully linked layer to retrieve bi-temporal characteristics. It comprises a convolutional layer, a maximum pooling layer, and four Residual Blocks (ResBlock). Each residual unit consists of three 3×3 convolutional layers. After being fused, the features are fed into the ReLU layer by mixing each attribute's top and bottom levels. The feature extraction unit retrieves four scales, and the MSA block applies position attention to their properties. The position attention module is used for upsampling and connecting additional ResBlock feature extraction results with low-level features to obtain local features at different scales. The following is a description of the MSA procedure:

$$Fi = Con\_cat(\sum_{n=1}^{i} Res\ Block - n), i = 1,2,3,4 \quad (1)$$

The core characteristics are fused with the different scale attention maps produced by the position attention module after the 1×1 Convolution. The feature maps in this have sizes of $\left[1,\frac{1}{2},\frac{1}{4},\frac{1}{8}\right]H'\times W'$. The multiscale characteristics acquired by the feature extraction and MSA components are then extrapolated to the actual picture dimension by the estimation component using bilinear interpolation. It preserves the complex spatial relationships and properties of the feature vectors.

3.3. Echo State Network for Caption Generation

ESN receives the extracted picture features and preprocessed image captions to generate appropriate captions. This network provides an organized representation of the input captions by analyzing the preprocessed captions to identify the temporal correlations in the text. Its reservoir dynamics and recurrent connections allow it to identify semantic links and temporal patterns in the text data. Finally, we included the fully connected layer at the end of the ESN to generate the relevant captions based on the relationship between the visual and text features.

An ESN is a particular kind of Recurrent Neural Network (RNN) where the RNN hidden layer is replaced with a reservoir. Unlike RNN models, the ESN model increases computational efficiency by altering the output weight matrix by applying a linear learning algorithm on a fixed recurrent neural network that serves as a reservoir. Its components are a reservoir, an output, and an input layer. The quantity of input, output, and reservoir neurons is denoted by the letters L, N, and K, respectively.

Let $u=(n)$ stand for the external input, $y=(n)$ for the output vector, and $x=(n)$ for the reservoir state. W_{out}, W, W_{back} and W_{in} represent the weight matrices of the output, reservoir, feedback, and input, respectively. The matrices' sizes are given in descending order by $L\times(K+N)$, $N\times L$, $N\times$B, and $N\times K$. The reservoir states were updated throughout the ESN training phase before learning the weight matrix W_{out} for the reservoir-to-output layer. The following is a definition of the ESN system dynamics.

$$x(t+1) = f(W_{in}u(t+1) + Wx(t) + W_{back}y(t)) \quad (2)$$

$$y(t+1) = f^{out}(W_{out}x(t+1)) \quad (3)$$

Here, $f(.)$ indicates a sigmoid function within the reservoir state, and f^{out} stands for the output's activation function.

The core idea of ESN is to use the reservoir as a temporal feature extraction tool. Its large dimensionality makes it possible to capture the temporal structure of the input features. These are handled in turn by several linked reservoirs. Each reservoir transforms the input features before being supplied into the next reservoir in the sequence. The final reservoir produces a result by performing a linear regression on the retrieved characteristics obtained after altering the input. Lastly, we link the reservoir output of the ESN to a fully linked layer, which serves as the caption-generating layer. The model can create captions word by word since the fully linked layer has a softmax activation that creates a probability spectrum over the vocabulary of possible phrases. Moreover, the hyperparameters are tuned using the TSA, described in the following subsection, to enhance the network's performance.

3.3.1. Hyperparameter Tuning Using Tunicate Swarm Algorithm

TSA for optimal hyperparameter tuning of the ESN model. The social nature of tunicates searching for food serves as an inspiration for TSA. The marine creature uses water jet and swarm intelligence to locate prey during hunting. The atrial siphons on all the tunicates allowed them to quickly expel the seawater they had breathed, generating a form of propelling energy that allowed them to move forward. Additionally, the tunicate showed SI after it could exchange search parameters on the meal's location. The tunicate must meet the following conditions to develop a mathematical model for its propellant jet mechanism. They are: 1) Prevent all search agents from colliding; 2) Ensure that all agents proceed in the direction of the fit person; and 3) Form search agent alliances in the area surrounding the fittest person.

The algorithm below was used to determine each search agent's unique location in order to avoid conflicts between them:

$$\vec{A} = \frac{\overrightarrow{GR}}{\overrightarrow{MP}} \quad (4)$$

$$\overrightarrow{GR} = c_2 + c_3 - \vec{F} \quad (5)$$

$$\overrightarrow{FS} = 2.c_1 \quad (6)$$

Here, $c1$, $c2$, and $c3$ represent three random integers within [0, 1] and $\vec{A}$ symbolize the vector utilized to locate the current position of all the agents. Gravity is denoted by $\overrightarrow{GR}$ and $\overrightarrow{FS}$ denotes the water movement in the deep sea. $\overrightarrow{MP}$ shows the relationship power among the searching agents as the vector value in the following manner:

$$\overrightarrow{MP} = P1min_{max_{min}} \quad (7)$$

Here, P_{max} and P_{min} denote the secondary and primary speeds in Equation (7) that enable the search agent to create social contact, and P_{max} and P_{min} are set to 1 and 4.

Once the conflicts between neighboring search agents have been resolved, each one moves in the direction of the neighbor with the highest Fitness Values (FV), which is presented as follows,

$$\overrightarrow{PDV} = \left|\overrightarrow{X_{best}} - r_{rand} \cdot \overrightarrow{X(t)}\right| \quad (8)$$

The food at the location of the current optimum individual is denoted as $\overrightarrow{X_{best}}$ in Equation (8) together with a $\overrightarrow{PDV}$ vector representing the spatial distance between the selected prey and tunicates, r_{rand} indicates an arbitrary number between zero and one, and $\overrightarrow{X(t)}$ is the precise position of the present search agent at the iterations.

The area was assessed to develop the search agent and conduct enough local investigation of nearby fittest individuals to identify the best solution for the present iteration:

$$X(t) = \begin{cases} \overrightarrow{X_{\text{best}}} - \vec{A} \cdot \overrightarrow{\text{PDV}}, if\ \text{r}_{\text{rand}} < 0.5 \\ \overrightarrow{X_{\text{best}}} + \vec{A} \cdot \overrightarrow{\text{PDV}}, if\ \text{r}_{\text{rand}} \geq 0.5 \end{cases} \quad (9)$$

During each iteration, every search agent investigates the area around the fittest person $\overrightarrow{X_{best}}$ and allocates the results $X(t)$ to improve the location.

The swarming behaviour of the tunicates allows the seeking agents to share location information. The current search agents' locations can influence this procedure, which can be completed by improving their locations. It can be achieved by the fittest person and the upgraded location by the previous person employing the swarm act:

$$X_i(\vec{t}+1) = \rightarrow \begin{cases} \dfrac{X_i(\vec{t}) + X_{i-1}(\vec{t}+1)}{2 + c_1} & if\ i > 1 \\ X_i(\vec{t}) & if\ i > 1 \end{cases} \quad (10)$$

In this case, i=1, .., N, the size of the population is denoted by N, $X_i(\vec{t}+1)$ indicates the location of the search agents that are now active, and $X_{i\text{-}1}(\vec{t}+1)$ is the location of the search agents from the previous iteration. Finally, the TSA algorithm selects the most optimal value for the ESN's hyperparameters at the end of the iteration. The settings consist of 100 epochs, 1e-06 weight decay rate, 0.001 learning rate, 8 batch size, 0.8 momentum, and 0.9 dropout rate.

4. Experimental Results

Execution of the suggested method and a summary of the study results for the automatic generation of captions for medical images are presented in this part. We contrasted the presented method with other

currently used medical image caption-generating techniques to evaluate the program's efficacy. The evaluation is conducted using two publically available datasets regarding various performance metrics, and Table 1 displays the setup for the experiment.

Table 1. Experimental setup.

Project	Environment
Operating system	Windows 10
Memory	16GB
CPU	i3-7100U
Language	Python 3.7
Framework	Keras

4.1. Dataset Description

The following four datasets served as the subjects of our experiments.

- **PEIR Gross dataset:** the PEIR-Gross dataset serves as a public medical picture resource for medical teaching. Every image in this dataset has a resolution of 528×792. It is composed of 7442 teaching images that are available to the public and are arranged into 21 predetermined categories. This dataset's entire number of image captions has a vocabulary size 4,452. Unlike a report, it features single-sentence captions, which sets it apart from the IU-XRay dataset. A caption for each image typically consists of 12 words.
- **IU X-ray dataset:** the open access biomedical image search engine provides public access to the IU X-ray dataset. The X-rays in this collection were in Portable Network Graphics (PNG) file format, with 512×624 resolution and 24 bits of depth. The radiology report for the 7,470 frontal and lateral chest X-rays is divided into four sections: the indication, comparison, findings, and impression. The intended medical reports for this study were created using physicians' findings.
- **MIMIC-CXR dataset:** this vast publicly accessible dataset includes 227,827 reports with 377,110 X-ray images of 65,379 patients. This dataset includes pictures with many views. In our research, we have utilized only frontal and lateral view images. A label tool called CheXpert is used to label the dataset for 14 typical chest radiography observations taken from the free-text radiology reports.
- **ROCO dataset:** the images in the Radiology Objects in Context (ROCO) dataset were sourced from PubMedCentral, an open-access archive of medical literature. It includes more than 81,000 radiology images from various types of imaging such as mammography, PET, fluoroscopy, X-ray, ultrasound, CT, and MRI. Every image in the ROCO collection has a unified medical language systems concept unique identifiers, national library of medicine's, semantic types, pertinent keywords, and the description.

4.2. Performance Metrics

The evaluation metrics used in the study are Metric for Evaluation of Translation with Explicit Ordering (METEOR), Metric for Consensus-based Image Description Evaluation (CIDEr), Recall-Oriented Understudy for Gisting Evaluation-Longest Common Subsequence (ROUGE-L), and Bilingual Evaluation Understudy (BLEU), which are word overlap measures.

METEOR and BLEU assess the translation outcomes. It comprehensively evaluates factors like memory, unigram precision, and consistency across the produced and reference texts. With a focus on encapsulating the overall meaning, ROUGE-L assesses the longest common subsequence among the created text and reference text. A statistic called CIDEr was created expressly to assess picture descriptions. It emphasizes the significance of capturing various facets of the image content in the captions by considering the consensus among several human annotators regarding the calibre of the automatically generated captions.

The BLEU score has a possible range of 0.00 to 1.00. Several BLEU ratings (BLEU-1 to BLEU-4) thoroughly assess the image captioning algorithms' effectiveness at various n-gram levels. The mathematical formulation for BLEU is,

$$BLEU = BP.exp(\sum_{n=1}^{N} w_n \log p_n) \quad (11)$$

$$w_n = \frac{1}{n} \quad (12)$$

$$BP = \begin{cases} 1 & if\ c > r \\ e^{(1-r/c)} & if\ c \leq r \end{cases} \quad (13)$$

$$P_n = \frac{\sum_{C\in\{candidates\}}\sum_{n\text{-}gram\in C} count_{clip}(n\_gram)}{\sum_{C'\in\{candidates\}}\sum_{n\text{-}gram'\in C'} count(n\_gram')} \quad (14)$$

$$count_{clip} = min(count, max\_ref\_count) \quad (15)$$

Here, $n\in\{1, 2, 3, 4\}$, the candidate and reference translation's length are denoted as c and r correspondingly, and w_n signifies the weight of n_gram. The amount of occurrences of n_ram in the reference and candidate is represented by (n_gram) and (n_gram') respectively. The term "count" denotes the quantity of instances of "n_gram" in the candidate, whereas the term "max-ref-count" denotes the highest amount of "n_gram" in the reference.

The mathematical formulation for *METEOR* is,

$$METEOR = (1 - pen) \times F_{means} \quad (16)$$

$$pen = \frac{\#chunks}{m} \quad (17)$$

$$F_{means} = \frac{PR}{\alpha P + (1-\alpha)R} \quad (18)$$

$$P = \frac{m}{c} \quad (19)$$

$$R = \frac{m}{r} \quad (20)$$

In this case, $\#chunks$ indicates the total amount of

chunks (the measure of grouping of close-matched 1-tuples in the reference and candidate translations), and the penalty factor is denoted by pen. The total amount of pairs that can be compatible in the candidate translation is indicated by m, the candidate translation's length is represented by c, the reference translation's length is indicated by r, and the controllable parameter is α.

Then, the mathematical formulation for *ROUGE-L* is,

$$ROUGE - L = \frac{(1-\beta^2)R_{lcs}\,P_{lsc}}{R_{lcs} + \beta^2\,P_{lsc}} \quad (21)$$

$$R_{lsc} = \frac{LSC(X,Y)}{m} \quad (22)$$

$$P_{lsc} = \frac{LSC(X,Y)}{n} \quad (23)$$

Here, the candidate translation is represented by X, the reference translation is Y, the length of the longest common subsequence among the reference and the candidate translation is represented by (X, Y) and the candidate and reference translation's length are represented by n and m correspondingly.

The mathematical formulation for *CIDEr* is,

$$CIDEr = \frac{CIDEr_{n-gram}}{CIDEr_{ref}} \quad (24)$$

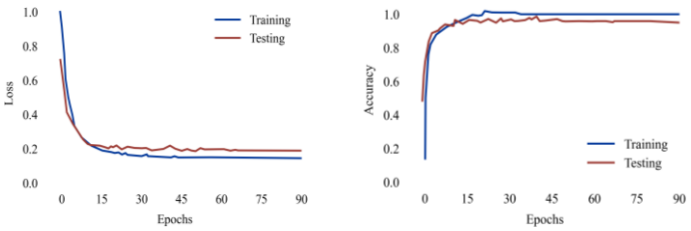
The *n-gram* based *CIDEr* score is represented as $CIDEr_{n\text{-}gram}$, and the normalization factor based on reference captions is called $CIDEr_{ref}$.

4.3. Training and Testing Evaluation

To assess performance, the datasets are divided into two sections at random. 20% was reserved for testing and the remaining 80% for training. The dataset wise training and testing samples are given in Table 2. The testing dataset is used to evaluate a model's performance after it has been trained on the training dataset. The TSA algorithm optimizes the following hyperparameters for the training of the proposed framework: 0.8 momentum, 0.9 dropout rate with L2 regularization, 1e-06 weight decay rate, and 0.001 learning rate. The batch size is set at 8. The model was trained throughout over 100 epochs to achieve optimal learning consistency. The accuracy statistic displays the proportion of accurate predictions made by the network once training is complete. It is illustrated by the accuracy curve, which displays the evolution of the proposed network. Figures 2, 3, 4, and 5 show the loss and accuracy during testing and training for the PEIR-Gross, IU X-ray, MIMIC CXR and ROCO datasets. After 15 epochs, the testing and training lines in the images converged, indicating that the network had stabilized.

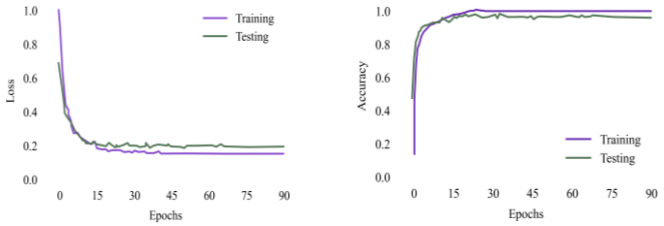
Table 2. Training and testing samples.

Datasets	Total samples	Training (80%)	Testing (20%)
PEIR-GROSS	7,442	5,953	1,489
IU X-ray	7,470	5,976	1,494
MIMIC CXR	250,022	200,018	50,004
ROCO	81,118	64,894	16,224



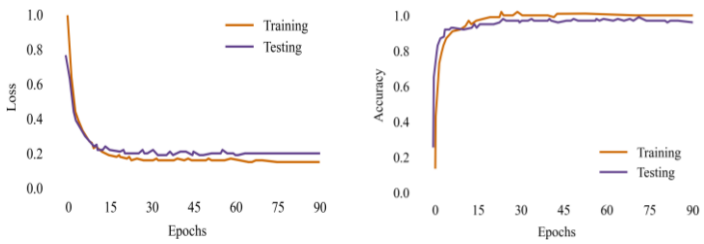
a) Loss. b) Accuracy.

Figure 2. Loss and accuracy graph for testing and training on the PEIR-GROSS dataset.



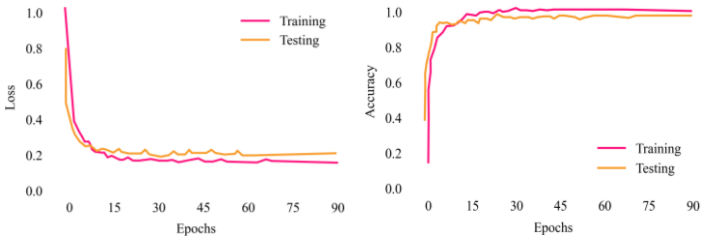
a) Loss. b) Accuracy.

Figure 3. Loss and accuracy graph for testing and training on IU X-ray dataset.



a) Loss. b) Accuracy.

Figure 4. Loss and accuracy graph for testing and training on the MIMIC CXR dataset.



a) Loss. b) Accuracy.

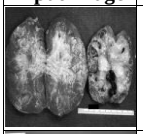
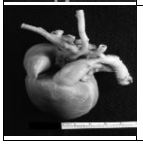
Figure 5. Loss and accuracy graph for testing and training on the ROCO dataset.

The testing accuracy and loss images also show that the recommended method does not overfit the training set. Furthermore, it is found that as the number of epochs increases, loss value in testing and training also decreases. In a similar vein, accuracy increases as the number of epochs increases.

4.4. Performance Analysis on PEIR GROSS Dataset

This subsection presents the outcomes of the suggested approach's performance evaluation using the PEIR-Gross dataset. We begin by analyzing the suggested approach's qualitative outcomes. Table 3 displays the model predictions, the ground truth expert label, and five original medical pictures from the PEIR-Gross dataset. We discovered that most of the anticipated and real captions for the image match exactly. Thus, we were given a maximum BLEU score for these captions. It suggests that the suggested framework uses both global and local information that is extracted by MASNet to find accurate captions. By integrating attention mechanisms, MASNet can concentrate on pertinent regions of interest within medical images. By doing this, the network's contextual awareness is improved, making it possible to extract features more closely associated with the image's descriptive or diagnostic qualities-additionally, the ESN's temporal processing skills aid in creating more accurate captions.

Table 3. Qualitative results on the PEIR-Gross dataset.

Input image	Ground truth	Predicted
	Urinary: Kidney: Staghorn calculus: gross good example opposite kidney appears to show compensatory hypertrophy.	Urinary system: Kidneys: Notable presence of a staghorn calculus is evident. Conversely, the kidney on the opposite side demonstrates compensatory hypertrophy.
	Urinary: Kidney: Goodpastures disease: gross granular kidney not at all typical for this disease but illustrates that gross is not typical in this disease.	Urinary: Kidney: Goodpastures disease: Uncommonly granular kidney presentation, highlighting atypical gross features in this condition.
	Urinary: Kidney: Acute tubular necrosis: Gross good example swollen cortex secondary to body burn.	Urinary: Kidney: Acute tubular necrosis: swollen cortex observed, likely due to body burn.
	CARDIOVASCULAR: Aorta: Coarctation: Gross hypoplastic aortic arch infantile coarctation well demonstrated great vessels.	Cardiovascular: Aorta: coarctation: Clear visualization of a hypoplastic aortic arch in infantile coarctation, showing the major blood vessels.

We examined the quantitative data and contrasted it with the methods already used to assess the suggested approach's performance even further. The outcomes of the suggested method are shown in Table 4, together with those of other methods. We can see from the table that the suggested method yields better results across the board for all performance indicators. When the results were examined immediately, it proved to be strong at continuously generating text and produced a report of excellent quality. The suggested framework effectively combines the attention mechanism of MASNet with the temporal processing of ESN, outperforming previous methods in terms of performance.

Table 4. Comparison of the presented method on the PEIR-Gross dataset.

Techniques	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METOR	CIDEr	ROUGE-L
Adaptive Multimodal Attention Network (AMANet) [42]	0.200	0.144	0.112	0.093	0.247	0.700	0.100
Medskip [22]	0.399	0.278	0.209	0.148	0.176	-	0.414
CNN-KQF [53]	-	-	-	0.142	0.330	-	0.156
YOLOv4-LSTM [25]	-	-	-	0.717	0.785	-	0.652
CNN-LSTM [10]	0.300	0.218	0.165	0.113	0.149	0.329	0.279
BioMedBLIP [20]	0.248	-	-	-	-	-	-
BioMEDGPT [49]	-	-	-	-	0.147	0.258	0.240
Proposed	0.889	0.857	0.836	0.812	0.886	0.836	0.798

In contrast to alternative methods, AMAnet yields inferior outcomes. Only 0.093, 0.247, 0.100, and 0.700 were obtained for BLEU-4, METOR, ROUGE-L, and CIDEr, respectively, due to the significant amount of data needed to learn the correlations among characteristics and produce accurate captions successfully. To overcome this problem, we used various data augmentation techniques in our proposed work to boost the training data, which improves the network's performance and yielded metrics of 0.812 for BLEU-4, 0.886 for METOR, 0.798 for ROUGE-L, and 0.836 for CIDEr.

In addition, BioMedBLIP and BioMEDGPT yield lower values than our method. Due to their heavy reliance on large-scale pretrained language models, these systems may not adapt as effectively to the unique features of medical imaging data. Furthermore, their systems frequently fail to integrate temporal context and fine-grained visual data effectively, producing captions that are less accurate and occasionally generic. Compared to these methods, our suggested model achieves a high BLEU-1 score of 0.889 and a CIDEr score of 0.836. By improving the extraction of intricate multi-scale characteristics, MASNet technically strengthens the model's resistance to noise and fluctuations in medical pictures. With fewer parameters and faster training, ESN's fixed recurrent reservoir facilitates effective temporal modeling, reducing overfitting and enhancing generalization to fresh data. In comparison to transformer-based models, this combination makes our model computationally efficient and enables strong performance on PEIR-Gross dataset.

4.5. Performance Analysis in IU X-Ray Dataset

We examined the IU X-ray dataset performance to further confirm the suggested method's effectiveness. We examined the reports produced by the suggested model, as seen in Table 5, to demonstrate the effectiveness of the offered method. The predicted caption reflects the generated report of the image, while the ground truth is the original report of the image. According to Table 5, the report produced using our approach is more representative of people's verbal and written patterns and more consistent with the original. It produces fair and understandable reports for medical diagnosis by considering irregular phrases as templates and incorporating the semantic consistency of medical words. In conclusion, our reports are correct and fluent and follow the conventions of medical professionals when writing medical diagnoses.

Table 6 compiles the experimental results for the IU X-ray dataset. We evaluated our outcome with CIDEr, ROUGE-L, METOR, and BLEU-4 measures. The proposed approach's ability to generate captions automatically is demonstrated by comparing it with five existing approaches. The results demonstrated that the proposed framework achieves significant values for each evaluation indicator, such as 0.812 for CIDEr metrics, 0.791 for ROUGE-L, 0.784 for METOR, and 0.654 for BLEU-4.

Table 5. Qualitative results on IU X-ray dataset.

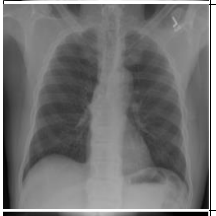
Input image	Ground truth	Predicted
	No acute cardiopulmonary abnormality. The cardiomediastinal silhouette and pulmonary vasculature are within normal limits. There is no pneumothorax or pleural effusion. There are no focal areas of consolidation. Cholecystectomy clips are present. Small T-spine osteophytes. There is biapical pleural thickening, unchanged from prior. Mildly hyperexpanded lungs.	No acute cardiopulmonary abnormalities are observed. The cardiomediastinal silhouette and pulmonary vasculature appear normal. No evidence of pneumothorax or pleural effusion is found. Focal areas of consolidation are absent. Cholecystectomy clips are identified. Small osteophytes are noted in the T-spine. Biapical pleural thickening remains unchanged from previous imaging. Mild hyperexpansion of the lungs is observed.
	No acute cardiopulmonary findings. Cardiomediastinal silhouette and pulmonary vasculature are within normal limits. Lungs are clear. No pneumothorax or pleural effusion. Evidence of prior granulomatous disease. No acute osseous findings.	No acute cardiopulmonary abnormalities detected. Cardiomediastinal silhouette and pulmonary vasculature appear normal. Lungs are clear without signs of pneumothorax or pleural effusion. Some evidence of past granulomatous disease observed. No acute bone abnormalities identified.
	No acute disease. The heart is normal in size. The mediastinum is unremarkable. The lungs are clear.	No acute issues found. Heart size normal. Mediastinum and lungs appear clear.
	No acute cardiopulmonary findings Heart size within normal limits. No focal alveolar consolidation, no definite pleural effusion seen. No typical findings of pulmonary edema. No pneumothorax.	There are no immediate heart or lung concerns. The heart appears normal in size. No concentrated alveolar consolidation or clear signs of pleural effusion are evident. There are no typical signs of pulmonary edema or pneumothorax.

Table 6. Comparison of proposed method on IU X-ray dataset.

Techniques	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METOR	CIDEr	ROUGE-L
X-Transformer [24]	0.428	0.265	0.178	0.119	0.198	0.458	0.337
Cross encoder-decoder [13]	0.506	0.319	0.220	0.160	0.192	-	0.380
ASGMD network [34]	0.489	0.326	0.232	0.173	0.206	-	0.397
CNN-LSTM [33]	0.580	0.342	0.263	0.155	-	-	-
Densenet-121 [37]	0.503	0.333	0.236	0.175	-	0.331	0.360
BioGPT [21]	0.668	0.625	0.569	0.481	0.189	0.415	0.774
XRaySwinGen-GPT-2 [17]	0.377	0.239	0.168	0.124	0.322	-	0.300
Proposed	0.805	0.745	0.698	0.654	0.784	0.812	0.791

In contrast, according to the table, the efficiency of recent transformer-based and hybrid models, such as BioGPT and XRaySwinGen-GPT-2, is average; BioGPT achieves a BLEU-1 of 0.6685 and a CIDEr of 0.4158. Our suggested approach clearly outperforms these, achieving a BLEU-1 of 0.805 and a CIDEr of 0.812, indicating gains in caption relevance and accuracy. This outcome demonstrates the benefit of integrating MASNet and ESN. Although transformer models like BioGPT use attention to capture global context, they are computationally demanding and need large amounts of data, and their global emphasis frequently degrades the representation of fine-grained local features. This is addressed by MASNet, which extracts intricate multi-scale local features that are essential for identifying minute patterns in medical images. Unlike conventional transformers, ESN effectively models temporal relationships with no training overhead. ESN's lightweight sequence modeling and MASNet's powerful spatial encoding combine to create a captioning framework that is both more accurate and computationally efficient. For medical imagine captioning tasks, this collaboration improves efficiency and feasibility by overcoming the inherent limits of pure attention-based models.

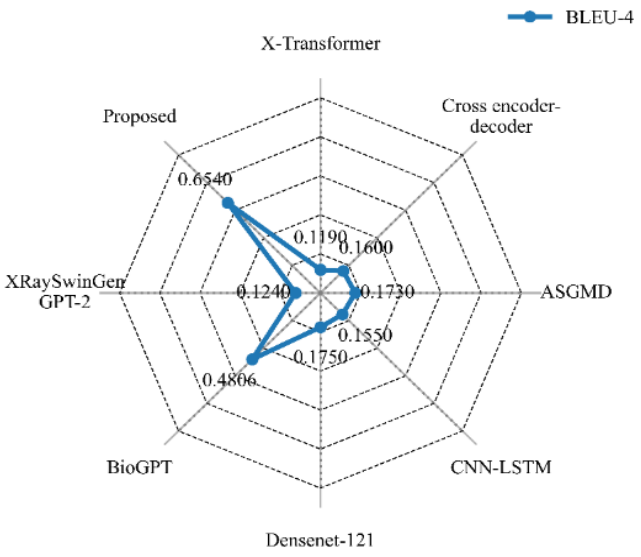


Figure 6. Graphic illustration of the BLEU-4 value on the IU X-ray dataset.

Furthermore, in contrast to alternative methods, the generated text's high BLEU score (Figure 6) shows that it covers more pertinent information in the reference text while avoiding superfluous repetition or verbosity. Additionally, the CIDEr score shows the Inverse Document Frequency (IDF) of every vocabulary word across the evaluated dataset. Conversely, the CIDEr measure gives greater weight to the essential but rare phrases that are taken into account by IDF. As a result, when it comes to writing medical reports, the CIDEr metric matters more. For this metric, the presented technique achieved 0.812. Therefore, it is highly recommended that the medical image captions be created using this method.

4.6. Performance Analysis in MIMIC-CXR Dataset

We conducted a qualitative analysis of the MIMIC-CXR dataset, seen in Table 7, to learn more about the reliability and clarity of the proposed caption generation. The suggested model exhibits higher sentence similarity and more detailed results when compared to the ground truth. This method shows that the suggested model is more effective at extracting visual information and converting it into text, similar to the vocabulary used by physicians. It shows that the suggested model uses both global and local aspects since it extracts osseous structure or aberrant sizes from the regional data. Additionally, ESNs are capable of handling the sequential nature of captioning with efficiency, resulting in the production of coherent and fluid sentences. At last, the predicted caption exhibits increased precision, encompassing crucial radiological discoveries in the X-ray picture. These findings demonstrate how well our suggested strategy can capture the crucial aspects and subtleties of the medical images while producing proper captions with high contextual relevance and descriptive richness.

Following the visual analysis, we evaluate the quantitative outcome of the suggested approach on MIMIC-CXR dataset. The findings are displayed in Table 8 and Figure 7. Our suggested method produced better outcomes than any other model. Our higher ROUGE-L and BLEU-n scores demonstrate the proposed method's ability to produce more cognitively fluent phrases. The level of similarity between a produced description and the source text is directly compared using BLEU scores. The BLEU-n scores in our proposed study are relatively high compared to previous medical image caption-generating tasks.

Table 7. Report generation outcomes on the MIMIC-CXR dataset.

Input image	Ground truth	Predicted
	A single portable AP semi-upright view of the chest was obtained. Heart is mildly enlarged. Calcifications are present in the aortic arch. Deviation of the trachea to the right is probably due to an enlarged thyroid. There are diffuse bilateral opacities with perihilar distribution and more prominent in the lower zone, consistent with moderate-to-severe pulmonary edema. Moderate bilateral pleural effusions are also present. The degree of pulmonary edema limits assessment for focal consolidation. There is no pneumothorax.	A single portable AP semi-upright chest view was obtained, revealing mild enlargement of the heart and calcifications in the aortic arch. Rightward tracheal deviation is likely due to an enlarged thyroid. Diffuse bilateral opacities are observed with a perihilar distribution, more prominent in the lower zones, consistent with moderate-to-severe pulmonary edema. Moderate bilateral pleural effusions are also present. The extent of pulmonary edema limits the evaluation for focal consolidation, and there is no evidence of pneumothorax.
	The PA catheter again extends to the right pulmonary artery within the mediastinal contours. Other monitoring and support devices remain unchanged. Continued enlargement of the cardiac silhouette without vascular congestion or pleural effusion or acute focal pneumonia.	The PA catheter extends into the right pulmonary artery, and other monitoring devices are unchanged. The cardiac silhouette continues to enlarge without vascular congestion, pleural effusion, or acute pneumonia.
	Two frontal images of the chest demonstrate a left basilar hazy opacity concerning for left lower lobe pneumonia. There is no pleural effusion or pneumothorax. There is some vascular crowding likely due to low lung volumes from poor inspiration. Heart size is normal.	Two frontal chest images show a hazy opacity in the left basilar region, raising concern for left lower lobe pneumonia. There is no pleural effusion or pneumothorax, but some vascular crowding is likely due to low lung volumes from poor inspiration. The heart size is normal.
	No comparison. Borderline size of the cardiac silhouette. No pleural effusions. Mild fluid overload but no overt pulmonary edema. Minimal increase in radiodensity at the bases of the right medial lung. The change should be radiographically monitored within 24 hours to exclude developing pneumonia.	No comparison. The cardiac silhouette is borderline in size, with no pleural effusions present. There is mild fluid overload but no overt pulmonary edema. A minimal increase in radiodensity is noted at the bases of the right medial lung, which should be monitored radiographically within 24 hours to rule out developing pneumonia.

Table 8. Comparison of the proposed approach on the MIMIC-CXR dataset.

Techniques	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METOR	CIDEr	ROUGE-L
CP-LIMHA [16]	0.362	0.227	0.155	0.113	0.142	-	0.283
ATN-CAN-CMA [23]	0.377	0.240	0.158	0.112	0.155	-	0.288
Knowledge Base and Multimodal Alignment Model (KBMMA) [43]	0.386	0.237	0.157	0.111	-	0.111	0.274
MKMIA [52]	0.399	0.242	0.158	0.109	-	-	0.275
Fully Transformer-based Encoder-Decoder framework (FTED) [46]	0.354	0.225	0.145	0.127	0.147	-	0.286
GIT-CXR [32]	0.403	0.286	0.215	0.168	0.369	-	0.312
Crossmodal Augmented Transformer (CAT) [35]	0.491	0.327	0.233	0.176	0.195	0.457	0.383
Proposed	0.768	0.621	0.522	0.486	0.537	0.699	0.678

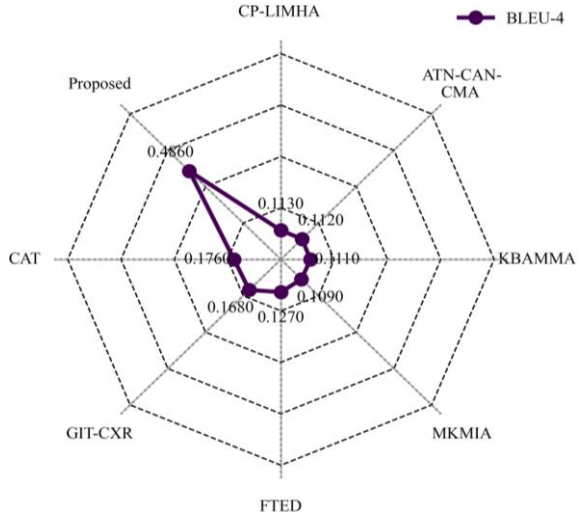


Figure 7. BLEU-4 comparison using the MIMIC CXR dataset.

Moreover, according to the table, transformer-based techniques such as FTED and CAT produce scores that are comparatively moderate; FTED displays a BLEU-1 of 0.354, while CAT raises it to 0.491. However, these results are considerably lower than our suggested method, which achieves a BLEU-1 of 0.768. The main drawbacks of pure transformer-based designs are highlighted by this performance disparity. Although FTED and CAT are good at using self-attention to capture global context, their scalability and practical utility are limited by their high computational cost and memory expense. More crucially, they miss fine-grained spatial details-like minor lesions or texture variations-that are crucial for medical picture description because they extensively rely on global attention mechanisms. As a result, the captioning performance is less accurate.

On the other hand, our method leverages the complimentary characteristics of MASNet and ESN by combining them and attain better results than existing techniques. For instance, our approach significantly raises the CIDEr score-52.1% higher on the MIMIC-CXR dataset. Our model's ability to preserve more critical details in the report generation is demonstrated by its higher CIDEr score (0.699). These findings show how well the suggested model could facilitate visual extraction of characteristics during training and offer

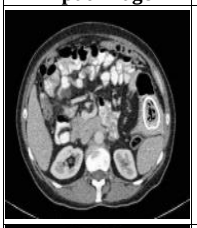
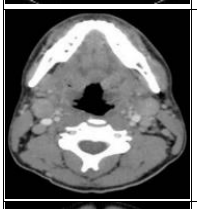
helpful radiological information. This result is achieved due to the effective feature extraction and caption generation capability of MASNet and ESN. It is possible for MASNet to comprehend both minute details and more comprehensive contextual information since it can gather features at various scales. Furthermore, the attention mechanism examines the extracted features and identifies the regions that contain the most information. By doing this, the network can focus on specific areas of the pictures essential for identifying variations or similarities. Moreover, an ESN's reservoir holds data about prior inputs, which enables the model to produce contextually appropriate captions using previously stored visual data. Every word produced while creating captions can affect the word after it. This sequential decision-making process can be effectively handled by ESNs, preserving relevance and coherence in the resulting content. Due to the advantages of these methods, the generated captions' word choice and sequence resemble the human-written descriptions, indicating that the descriptions of the medical images are more accurate and relevant.

4.7. Performance Analysis in ROCO Dataset

To qualitatively evaluate the efficiency of the suggested framework, we evaluate the outputs from the ROCO dataset, including the radiology report produced by our model, the expert-annotated ground truth report, and the input medical image. Table 9 shows that the projected report has a high degree of semantic alignment with the ground truth, accurately representing clinical descriptions (such as "oval density," "wall thickening") as well as anatomical characteristics (such as "small bowel," "left flank").

The primary reason for this great precision is the MASNet and ESN components' synergistic combination. A lightweight, attention-guided CNN called MASNet is excellent at identifying regional texture differences and fine-grained spatial characteristics that are crucial in medical imaging. Even in areas with poor contrast or anatomic complexity, its multi-scale aggregation approach guarantees reliable extraction of context-aware visual characteristics. ESN, on the other hand, ensures logical narrative flow and clinically sound perception by modeling temporal and syntactic relationships during report production by utilizing its dynamic reservoir computing architecture. Moreover, the produced reports' sentence structures reflect the rules of radiological reporting: accurate anatomic localization is presented first, followed by thorough clinical details, and, if appropriate, diagnostic observations at the end. The results improve clinical readability and interpretability by exhibiting accurate syntactic usage, logical sentence progression, and appropriate use of domain-specific vocabulary. To sum up, the MASNet-ESN framework shows a good ability to produce radiology reports that are both clinically correct and linguistically structured, confirming its potential for incorporation into automated diagnostic documentation systems in the real world.

Table 9. Qualitative results on the ROCO dataset.

Input image	Ground truth	Predicted
	CT scan showing oval calcification in a segment of the small bowel in the left flank situated within a diverticulum with thickening of the small bowel wall at that level.	CT scan showing oval density in a segment of the small bowel in the left flank within a diverticulum, with mild wall thickening at that level.
	CT scan of head and neck. Note cervical lymphadenopathy.	CT scan of head and neck shows cervical lymphadenopathy.
	CT scan demonstrating the presence of air in the left lobe of the liver extending beyond 2 Â cm from the liver capsule.	CT scan demonstrating air in the left lobe of the liver extending more than 2 cm from the liver capsule.
	Case 1. CECT demonstrates a splenic laceration extending to the splenic surface from a focal hypodense cyst-like lesion. There is a large perisplenic haematoma and abdominal free fluid	Case 1. CECT demonstrates a splenic laceration reaching the splenic surface from a hypodense cyst-like lesion, with perisplenic haematoma and abdominal free fluid.

The performance evaluation of several image-to-text generation models on the ROCO dataset is shown in Table 10 using common assessment metrics including BLEU (1-4), CIDEr, ROUGE-L, and METEOR. The BLIP model outperformed the other models in terms of BLEU scores (BLEU-1: 0.7959, BLEU-4: 0.7300) and ROUGE-L (0.8405), but its METEOR score (0.6101) and lack of CIDEr score demonstrate its inadequate semantic relevance and coverage. In a similar vein MSMedCap and VIT (SWIN)+BIOMEDBERT did not perform well; MSMedCap's ROUGE-L (0.154) and BLEU were particularly low, suggesting poor textual alignment with ground truth. The attention-based encoder-decoder models and Swin-S+VLT demonstrated modest gains, but they continued to have poor METEOR and CIDEr scores, which indicated insufficient contextual and semantic comprehension. These drawbacks arise mostly from their incapacity to properly utilize global-local interactions in medical imagery and textual properties, their poor semantic modeling, and their absence of strong attention mechanisms.

Table 10. Comparison of the proposed approach on the ROCO dataset.

Techniques	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METOR	CIDEr	ROUGE-L
Bootstrapping Language-Image Pre-training (BLIP) model [1]	0.7959	0.7714	0.7486	0.7300	0.6101	-	0.8405
VIT (SWIN)+BIOMEDBERT [18]	0.3348	-	-	-	-	-	-
MSMedCap [51]	0.1089	0.0481	0.0231	-	0.626	0.575	0.154
Swin-S+VLT [40]	0.491	0.452	0.436	0.420	0.282	0.3535	0.455
Attention-based encoder-decoder model [3]	0.4661	0.3297	0.2363	0.1861	-	-	-
Proposed	0.895	0.871	0.827	0.806	0.831	0.846	0.873

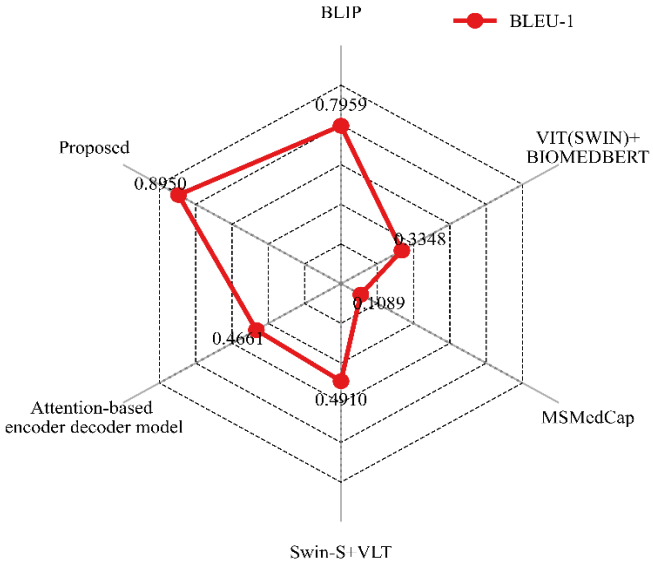


Figure 8. Comparison of BLEU-1 on the ROCO dataset.

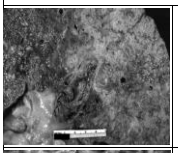
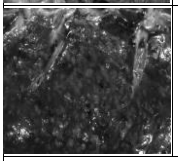
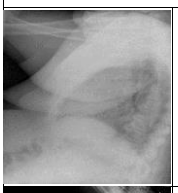
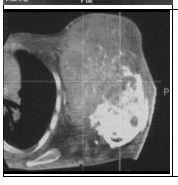
On the other hand, the suggested MASNET-ESN model performs noticeably better than all baselines, taking the top ranks on every metric (CIDEr: 0.846, METEOR: 0.831, ROUGE-L: 0.873, BLEU-4: 0.806, BLEU-1: 0.895). The suggested method's captions closely resemble the reference captions at the unigram level, exhibiting strong word-level accuracy and successful key term recognition in the medical images, as evidenced by the high BLEU-1 score (Figure 8). This better performance has been attributed by its temporal modeling based on ESNs and multi-scale attentive fusion, which improve long-term dependency learning and spatial feature integration. Fine-grained features are successfully captured by the application of multi-scale self-attention, and the ESN guarantees good temporal dynamics modeling without the need for complicated training. By resolving the shortcomings of earlier techniques, these benefits allow MASNET-ESN to provide medical reports that are more logical, contextually relevant, and semantically rich.

4.8. Failure Case Study

In Table 11, we performed a failed case study to further examine the effectiveness of the suggested approach by choosing unclear, poor-quality, or contradicting pictures from PEIR-Gross, IU X-ray, MIMIC CXR and ROCO datasets for in-depth analysis without any cleaning, normalization, or improvement. The model occasionally had trouble capturing fine details or particular clinical descriptions in these difficult cases, for example "spotty calcification" or unique lesion features. As a result, performance metrics were a little lower and the captions were less comprehensive. For instance, the model correctly depicted the tumor and the impact of surrounding tissues in a CT scan. However, it neglected to mention the existence of spotty calcifications, which are crucial diagnostic features. This absence slightly lowered the resulting report's precision and completeness. Nevertheless, in the majority of cases, the model demonstrated strong robustness by reliably identifying the primary pathogenic symptoms and providing clinically useful descriptions, even with these limitations.

For example, the model correctly identified all of these important details in an image that showed a "small right pneumothorax," "stable cardiomediastinal contours," and "properly positioned tubes," generating an appropriate and pertinent caption in spite of the image's complexity. This failure case analysis validates the model's general efficacy in actual medical picture captioning settings while highlighting opportunities for further improvement, namely in enhancing the identification and incorporation of fine-grained features in unclear or low-quality images.

Table 11. Failure case study.

Input image	Ground truth	Predicted	Performance metrics
PEIR-Gross dataset			
	Gross: Hepatobiliary: Liver: Biliary: Hepatoma: Gross macronodular cirrhosis with hepatoma very good.	Gross macronodular cirrhosis with hepatoma in the liver and biliary system showing hepatobiliary involvement	BLEU-4=0.712, CIDEr=0.740, ROUGE-L=0.702, METEOR=0.791
	Gross: Respiratory: Lung: Bronchopneumonia: Gross close-up good example of confluent bronchopneumonia.	Close-up of gross lung showing bronchopneumonia with confluent areas representing respiratory infection	BLEU-4=0.680, CIDEr=0.717, ROUGE-L=0.691, METEOR=0.753
IU X-ray dataset			
	Cardiomegaly with low lung volumes are clear. A XXXXXXXX lung volumes. Lungs are clear without focal airspace. No pleural effusions or pneumothoraces identified. Cardiomegaly. Degenerative changes observed in the spine.	Lungs are clear without focal airspace disease no pleural effusions or pneumothoraces cardiomegaly degenerative changes in the spine.	BLEU-4=0.568, CIDEr=0.705, ROUGE-L=0.730, METEOR=0.683
	Minimal perihilar opacity possibly representing atypical pneumonia. The heart size is at the upper limits of normal. Pulmonary XXXX and mediastinum are at normal limits. No pleural effusion or pneumothorax identified. Mild streaky perihilar opacity present without confluent airspace opacity to suggest bacterial pneumonia.	Mild perihilar opacity observed, which may indicate pneumonia. Heart size appears near normal limits. Pulmonary vasculature and mediastinum are unremarkable. No pleural effusion or pneumothorax is seen. Streaky opacity noted without evidence of bacterial pneumonia.	BLEU-4=0.589, CIDEr=0.694, ROUGE-L=0.700, METEOR=0.704
MIMIC CXR dataset			
	Small right pneumothorax stables since six hours prior. Right Internal Jugular (IJ) catheter tip in low Superior Vena Cava (SVC)/cavoatrial junction. Persistent low lung volumes. Stable cardiomediastinal contours. Moderate pulmonary edema unchanged. Possible small left pleural effusion. Endotracheal (ET) tube in standard position. Nasogastric (NG) tube tip not seen below diaphragm.	Small pneumothorax on the right side remains stable. Right internal jugular catheter tip located near the lower superior vena cava. Lung volumes appear low. Cardiomediastinal contours are stable moderate pulmonary edema is stable. Possible left pleural effusion present. et tube is in standard position ng tube tip is not visualized due to technique below the diaphragm.	BLEU-4=0.431, CIDEr 0.634, ROUGE-L=0.611, METOR=0.462
	Small right pneumothorax has increased compared to prior study performed the same day morning. NG tube tip out of view below diaphragm. ET tube in standard position. Right IJ catheter tip is in the lower SVC. low lung volumes. Mild-to-moderate pulmonary edema is stable. Cardiomediastinal contours are unchanged. Sternal wires are aligned.	Pneumothorax has increased compared to prior studies ng tube tip not visible below diaphragm. et tube positioned correctly ij catheter tip is in the lower svc there are low lung volumes mildtomoderate pulmonary edema mild to moderate and stable. Cardiomediastinal contours stable. Sternal wires noted.	BLEU-4 0.451, CIDEr=0.652, ROUGE-L=0.637, METOR=0.489
ROCO dataset			
	A 70-year-old gentleman presented with a three-month of localized left flank pain, mild to moderate in intensity. UHCT (Figure 1) revealed gross hydronephrosis and hydroureter with a 1 cm mid-ureteric calculus (right arrow). Incidentally, hyperdense bony deposits were noted in the L3 vertebral body (left arrow). Subsequent evaluation showed a raised Prostate-Specific Antigen (PSA), and histopathology of a prostate biopsy confirmed adenocarcinoma.	70 years' gentleman presented with a three month of localized left flank pain. It was mild to moderate in intensity. uhct showed gross hydronephrosis and hydroureter with a 1 cm mid-ureteric calculus. hyperdense bone deposits in the L3 vertebral body (left arrow). Further evaluation showed a raised PSA, and biopsy confirmed adenocarcinoma.	BLEU-4=0.734, CIDEr=0.751, ROUGE-L=0.780, METOR=0.746
	CT scan showing a moderately enhancing tumor destroying humeral head with a large osseous component extending into the surrounding muscles and soft tissues up to skin. Spotty calcification	scan showing a moderately enhancing tumor destroying humeral head with a large extra- osseous component extending and infiltrating the surrounding muscles and soft tissues up to skin.	BLEU-4= 0.718, CIDEr=0.742, ROUGE-L=0.763, METOR=0.723

4.9. Noise/Distortion Tests

We conducted experimental tests introducing two types of artificial distortions: motion blur and Gaussian noise, in order to thoroughly assess the robustness of our medical report generating algorithm under realistic imaging degradations (Table 12). Light (σ=0.01), medium (σ=0.05), and heavy (σ=0.10) Gaussian noise were added at progressively increasing severity levels to mimic sensor noise and low-dose imaging abnormalities that are frequently seen in clinical practice. Convolution was used to create motion blur using kernel sizes of 3, 5, and 7 pixels, which corresponded to light, moderate, and heavy patient motion artifacts during acquisition. The evaluation was carried out utilizing quantitative criteria such as CIDEr, ROUGE-L, METEOR, and BLEU-4, to thoroughly analyze caption quality across four benchmark datasets: PEIR-Gross, IU X-Ray, MIMIC-CXR, and ROCO.

According to experimental results, the degree of distortion causes a predicted but slow reduction in performance indicators. In particular, the model demonstrated a 12% decrease in BLEU-4 scores at the maximum noise level (σ=0.10) and a similar decline at the highest motion blur level (kernel=7). Importantly, our model showed remarkable robustness by maintaining high caption generation ability across all distortion kinds and datasets, even when significant image distortions were included. For example, the CIDEr score remained as high as 0.74 (PEIR-Gross), 0.71 (IU X-ray), 0.605 (MIMIC-CXR), and 0.75 (ROCO) under the highest Gaussian noise (σ=0.10), while the BLEU-4 scores remained above 0.70, 0.58, 0.41, and 0.715, respectively. These results show that the generated reports maintained semantic proficiency and clinical significance. Likewise, under severe motion blur (kernel size=7), the model obtained CIDEr scores of 0.76 (PEIR-Gross), 0.735 (IU X-ray), 0.63 (MIMIC-CXR), and 0.78 (ROCO) with only minor relative decreases (≈9-15%) from clean circumstances. Despite severe image deterioration, METEOR and ROUGE-L showed comparable patterns, indicating that language structure and semantic content were mostly maintained. These findings confirm that, although performance deteriorates as predicted with the degree of distortion, the model retains a significant level of captioning accuracy and generalizability, which is consistent with actual clinical settings where motion and noise aberrations are frequently inevitable.

Table 12. Noise tests on proposed approach.

Performance metric	None (clean)	Gaussian noise (σ=0.01)	Gaussian noise (σ=0.05)	Gaussian noise (σ=0.10)	Motion blur (kernel=3)	Motion blur (kernel=5)	Motion blur (kernel=7)
PEIR-Gross dataset							
BLEU-4	0.812	0.795	0.761	0.712	0.781	0.755	0.732
METOR	0.886	0.87	0.832	0.791	0.861	0.841	0.815
ROUGE-L	0.798	0.78	0.745	0.702	0.768	0.745	0.720
CIDEr	0.836	0.820	0.784	0.741	0.807	0.785	0.760
IU X-ray dataset							
BLEU-4	0.654	0.640	0.615	0.581	0.63	0.61	0.590
METOR	0.784	0.770	0.738	0.701	0.752	0.735	0.715
ROUGE-L	0.791	0.773	0.745	0.712	0.763	0.74	0.715
CIDEr	0.812	0.790	0.752	0.715	0.78	0.76	0.735
MIMIC-CXR dataset							
BLEU-4	0.486	0.471	0.445	0.411	0.455	0.44	0.425
METOR	0.537	0.521	0.49	0.455	0.508	0.49	0.475
ROUGE-L	0.678	0.660	0.625	0.580	0.648	0.63	0.611
CIDEr	0.699	0.682	0.648	0.605	0.670	0.65	0.630
ROCO dataset							
BLEU-4	0.806	0.788	0.755	0.715	0.775	0.755	0.735
METOR	0.831	0.815	0.785	0.745	0.802	0.78	0.760
ROUGE-L	0.873	0.855	0.82	0.778	0.841	0.815	0.791
CIDEr	0.846	0.830	0.795	0.750	0.820	0.811	0.782

5. Conclusions

Natural language captions for medical photographs convey the visual information contained in the photos. To assist medical practitioners in producing reports more precisely and quickly, we investigated how to automatically generate textual reports for medical photographs in this work. To address this, we developed a framework based on MASNet-ESN, which can examine an image's visual and semantic aspects and more successfully capture long-range semantics to generate lengthy sentences of excellent quality. We assessed the suggested approach's efficacy using quantitative and qualitative results on two medical datasets with radiology and pathology images. The proposed model performed better than the existing models, with 0.654 for BLEU-4, 0.784 for METOR, 0.791 for ROUGE-L, and 0.812 for CIDEr, according to the findings obtained using the IU X-RAY dataset. Similarly, it achieved 0.812 for BLEU-4, 0.886 for METOR, 0.798 for ROUGE-L, and 0.836 for CIDEr on the PEIR-Gross dataset, 0.806 for BLEU-4, 0.831 for METOR, 0.873 for ROUGE-L, and 0.846 for CIDEr, and 0.486 for BLEU-4, 0.537 for METOR, 0.678 for ROUGE-L, and 0.699 for CIDEr on the MIMIC CXR dataset. It illustrates the fantastic performance of the method and enables clinical decision support. This work is extended to various types of outdoor photographs and other medical image modalities in future research.

References

[1] Abed E. and Aguili T., "Automated Medical Image Captioning Using the BLIP Model: Enhancing Diagnostic Support with AI-Driven Language Generation," *Diyala Journal of Engineering Sciences*, vol. 18, no. 2, pp. 228-248, 2025. https://doi.org/10.24237/djes.2025.18215

[2] Ayesha H., Tariq M., and Israr S., "Computer Aided Deep Image Captioning for Medical Images," *Machines and Algorithms*, vol. 2, no. 1, pp. 1-16, 2023. https://knovell.org/MnA/index.php/ojs/article/view/36

[3] Beddiar D., Oussalah M., and Seppanen T., "Retrieved Generative Captioning for Medical

Images," *in Proceedings of the 20th International Conference on Content-based Multimedia Indexing*, Orleans, pp. 48-54, 2023. https://doi.org/10.1145/3617233.3617246

[4] Cao Y., Cui L., Zhang L., Yu F., and et al., "MMTN: Multimodal Memory Transformer Network for Image-Report Consistent Medical Report Generation," *in Proceedings of the AAAI Conference on Artificial Intelligence*, Washington (DC), pp. 277-285, 2023. https://doi.org/10.1609/aaai.v37i1.25100

[5] Cao Y., Ding H., Zhang Y., and Hei Y., "Radiology Report Generation Based on Adaptive Enhanced Fusion of Multi Features," *Computers in Biology and Medicine*, vol. 193, pp. 110494, 2025. https://doi.org/10.1016/j.compbiomed.2025.110494

[6] Chitteti C. and Madhavi K., "Taylor African Vulture Optimization Algorithm with Hybrid Deep Convolution Neural Network for Image Captioning System," *Multimedia Tools Applications*, vol. 83, pp. 1-19, 2024. https://doi.org/10.1007/s11042-023-18080-0

[7] Divya P., Sravani Y., Vishnu C., Mohan C., and Chen Y., "Memory Guided Transformer with Spatio-Semantic Visual Extractor for Medical Report Generation," *IEEE Journal of Biomedical and Health Informatics*, vol. 28, no. 5, pp. 3079-3089, 2024. https://doi.org/10.1109/JBHI.2024.3371894

[8] Elbedwehy S., Medhat T., Hamza T., and Alrahmawy M., "Enhanced Descriptive Captioning Model for Histopathological Patches," *Multimedia Tools Applications*, vol. 83, no. 12, pp. 36645-36664, 2024. https://doi.org/10.1007/s11042-023-15884-y

[9] Huang Z., Zhang X., and Zhang S., "KiUT: Knowledge-Injected U-Transformer for Radiology Report Generation," *in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Vancouver, pp. 19809-19818, 2023. https://doi.org/10.1109/CVPR52729.2023.01897

[10] Jing B., Xie P., and Xing E., "On the Automatic Generation of Medical Imaging Reports," *arXiv Preprint*, vol. arXiv:1711.08195v3, pp. 1-10, 2017. https://arxiv.org/abs/1711.08195v3

[11] Kim G., Oh B., Kim C., and Kim Y., "Convolutional Neural Network and Language Model-based Sequential CT Image Captioning for Intracerebral Hemorrhage," *Applied Sciences*, vol. 13, no. 17, pp. 1-13, 2023. https://doi.org/10.3390/app13179665

[12] Kong J., Oh B., Kim C., and Kim Y., "Sequential Brain CT Image Captioning Based on the Pre-Trained Classifiers and a Language Model," *Applied Sciences*, vol. 14, no. 3, pp. 1-15, 2024. https://doi.org/10.3390/app14031193

[13] Lee H., Cho H., Park J., Chae J., and Kim J., "Cross Encoder-Decoder Transformer with Global-Local Visual Extractor for Medical Image Captioning," *Sensors*, vol. 22, no. 4, pp. 1-13, 2022. https://doi.org/10.3390/s22041429

[14] Li M., Liu R., Wang F., Chang X., and Liang X., "Auxiliary Signal-Guided Knowledge Encoder-Decoder for Medical Report Generation," *World Wide Web*, vol. 26, no. 1, pp. 253-270, 2023. https://doi.org/10.1007/s11280-022-01013-6

[15] Lin Y., Lai K., and Chang W., "Skin Medical Image Captioning Using Multi-Label Classification and Siamese Network," *IEEE Access*, vol. 11, pp. 23447-23454, 2023. https://doi.org/10.1109/ACCESS.2023.3249462

[16] Lin Z., Zhang D., Shi D., Xu R., and et al., "Contrastive Pre-Training and Linear Interaction Attention-based Transformer for Universal Medical Reports Generation," *Journal of Biomedical* Informatics, vol. 138, pp. 104281, 2023. https://doi.org/10.1016/j.jbi.2023.104281

[17] Magalhaes G., Santos R., Vogado L., Paiva A., and Neto P., "XRaySwinGen: Automatic Medical Reporting for X-Ray Exams with Multimodal Model," *Heliyon*, vol. 10, no. 7, pp. 1-8, 2024. DOI: 10.1016/j.heliyon.2024.e27516

[18] Mayzura W., Sarno R., Suroto N., Supriyanto M., and Sihaj G., "Automatic Interpretation of Brain Medical Images Using Hierarchical Classification and Image Captioning Model," *IEEE Access*, vol. 13, pp. 84675-84688, 2025. https://doi.org/10.1109/ACCESS.2025.3560701

[19] Morampudi M., Gonthina N., Bhaskar N., and Reddy V., "Image Description Generator using Residual Neural Network and Long Short-Term Memory," *Computer Science Journal of Moldova*, vol. 31, no. 1, pp. 3-21, 2023. https://www.math.md/files/csjm/v31-n1/v31-n1-(pp3-21).pdf

[20] Naseem U., Thapa S., and Masood A., "Advancing Accuracy in Multimodal Medical Tasks Through Bootstrapped Language-Image Pretraining (BioMedBLIP): Performance Evaluation Study," *JMIR Medical Informatics*, vol. 12, no. 1, pp. 1-19, 2024. https://pubmed.ncbi.nlm.nih.gov/39102281/

[21] Ouis M. and Akhloufi M., "ChestBioX-Gen: Contextual Biomedical Report Generation from Chest X-Ray Images Using BioGPT and Co-Attention Mechanism," *Frontiers in Imaging*, vol. 3, pp. 1373420, 2024. https://doi.org/10.3389/fimag.2024.1373420

[22] Pahwa E., Mehta D., Kapadia S., Jain D., and Luthra A., "Medskip: Medical Report Generation Using Skip Connections and Integrated Attention," *in Proceedings of the IEEE/CVF International Conference on Computer Vision*,

Montreal, pp. 3409-3415, 2021. https://doi.org/10.1109/ICCVW54120.2021.00380

[23] Pan Y., Liu L., Yang X., Peng W., and Huang Q., "Chest Radiology Report Generation Based on Cross-Modal Multiscale Feature Fusion," *Journal of Radiation Research and Applied Sciences*, vol. 17, no. 1, pp. 100823, 2024. https://doi.org/10.1016/j.jrras.2024.100823

[24] Park H., Kim K., Park S., and Choi J., "Medical Image Captioning Model to Convey More Details: Methodological Comparison of Feature Difference Generation," *IEEE Access*, vol. 9, pp. 150560-150568, 2021. https://doi.org/10.1109/ACCESS.2021.3124564

[25] Ravinder P. and Srinivasan S., "Automated Medical Image Captioning with Soft Attention-Based LSTM Model Utilizing YOLOv4 Algorithm," *Journal of Computer Science*, vol. 20, no. 1, pp. 52-68, 2024. https://doi.org/10.3844/jcssp.2024.52.68

[26] Reddy P., Verma V., and Varma M., "Optimizing Medical Image Report Generation with Varied Attention Mechanisms," *in Proceedings of the 6th International Conference on Contemporary Computing and Informatics (IC3I)*, Gautam Buddha Nagar, pp. 2137-2143, 2023. https://doi.org/10.1109/IC3I59117.2023.10398149

[27] Revathi B. and Kowshalya A., "Automatic Image Captioning System Based on Augmentation and Ranking Mechanism," *Signal, Image Video Processing*, vol. 18, no. 1, pp. 265-274, 2024. https://doi.org/10.1007/s11760-023-02725-6

[28] Selivanov A., Rogov O., Chesakov D., Shelmanov A., and et al., "Medical Image Captioning via Generative Pretrained Transformers," *Scientific Reports*, vol. 13, no. 1, pp. 1-12, 2023. https://www.nature.com/articles/s41598-023-31223-5

[29] Shaik N. and Cherukuri T., "Gated Contextual Transformer Network for Multimodal Retinal Image Clinical Description Generation," *Image and Vision Computing*, vol. 143, pp. 104946, 2024. https://doi.org/10.1016/j.imavis.2024.104946

[30] Shao Z., Han J., Debattista K., and Pang Y., "DCMSTRD: End-to-End Dense Captioning via Multiscale Transformer Decoding," *IEEE Transactions on Multimedia*, vol. 26, pp. 7581-7593, 2024. https://doi.org/10.1109/TMM.2024.3369863

[31] Shentu J. and Al Moubayed N., "CXR-IRGen: An Integrated Vision and Language Model for the Generation of Clinically Accurate Chest X-Ray Image-Report Pairs," *in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, Waikoloa, pp. 5212-5221, 2024. https://doi.org/10.1109/WACV57701.2024.00513

[32] Sirbu I., Sirbu I., Bogojeska J., and Rebedea T., "GIT-CXR: End-to-End Transformer for Chest X-Ray Report Generation," *Information*, vol. 16, no. 7, pp. 1-27, 2025. https://doi.org/10.3390/info16070524

[33] Sirshar M., Paracha M., Akram M., Alghamdi N., and et al., "Attention Based Automated Radiology Report Generation Using CNN and LSTM," *PloS One*, vol. 17, no. 1, pp. 1-20, 2022. https://doi.org/10.1371/journal.pone.0262209

[34] Tan Y., Li C., Qin J., Xue Y., and Xiang X., "Medical Image Description Based on Multimodal Auxiliary Signals and Transformer," *International Journal of Intelligent Systems*, vol. 2024, pp. 1-12, 2024. https://doi.org/10.1155/2024/6680546

[35] Tang Y., Yuan Y., Tao F., and Tang M., "Cross-Modal Augmented Transformer for Automated Medical Report Generation," *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 13, pp. 33-48, 2025. https://doi.org/10.1109/JTEHM.2025.3536441

[36] Tiwary T. and Mahapatra R., "An Accurate Generation of Image Captions for Blind People Using Extended Convolutional Atom Neural Network," *Multimedia Tools Applications*, vol. 82, no. 3, pp. 3801-3830, 2023. https://doi.org/10.1007/s11042-022-13443-5

[37] Wang F., Liang X., Xu L., and Lin L., "Unifying Relational Sentence Generation and Retrieval for Medical Image Report Composition," *IEEE Transactions on Cybernetics*, vol. 52, no. 6, pp. 5015-5025, 2022. https://doi.org/10.1109/TCYB.2020.3026098

[38] Wang Y., Lin Z., Xu Z., Dong H., and et al., "Trust it or not: Confidence-Guided Automatic Radiology Report Generation," *Neurocomputing*, vol. 578, pp. 127374, 2024. https://doi.org/10.1016/j.neucom.2024.127374

[39] Xu D., Zhu H., Huang Y., Jin Z., and et al., "Vision-Knowledge Fusion Model for Multi-Domain Medical Report Generation," *Information Fusion*, vol. 97, pp. 101817, 2023. https://doi.org/10.1016/j.inffus.2023.101817

[40] Xu L., Liu B., Khan A., Fan L., and Wu X., "Multi-Modal Pre-Training for Medical Vision-Language Understanding and Generation: An Empirical Study with a New Benchmark," *arXiv Preprint*, vol. arXiv:2306.06494v2, pp. 1-18, 2023. https://arxiv.org/abs/2306.06494v2

[41] Xue Y., Tan Y., Tan L., Qin J., and Xiang X., "Generating Radiology Reports via Auxiliary Signal Guidance and a Memory-Driven Network," *Expert Systems with Applications*, vol. 237, pp. 121260, 2024. https://doi.org/10.1016/j.eswa.2023.121260

[42] Yang S., Niu J., Wu J., Wang Y., and et al., "Automatic Ultrasound Image Report Generation

with Adaptive Multimodal Attention Mechanism," *Neurocomputing*, vo. 427, pp. 40-49, 2021. https://doi.org/10.1016/j.neucom.2020.09.084

[43] Yang S., Wu X., Ge S., Zheng Z., and et al., "Radiology Report Generation with a Learned Knowledge Base and Multimodal Alignment," *Medical Image Analysis*, vol. 86, pp. 102798, 2023. https://doi.org/10.1016/j.media.2023.102798

[44] Yang X., Wang Y., Chen H., Li J., and Huang T., "Context-Aware Transformer for Image Captioning," *Neurocomputing*, vol. 549, pp. 126440, 2023. https://doi.org/10.1016/j.neucom.2023.126440

[45] Yang X., Yang Y., Wu J., Sun W., and et al., "CA-Captioner: A Novel Concentrated Attention for Image Captioning," *Expert Systems with Applications*, vol. 250, pp. 123847, 2024. https://doi.org/10.1016/j.eswa.2024.123847

[46] Zeiser F., Costa C., Ramos G., Maier A., and Righi R., "CheXReport: A Transformer-based Architecture to Generate Chest X-Ray Reports Suggestions," *Expert Systems with* Applications, vol. 255, pp. 124644, 2024. https://doi.org/10.1016/j.eswa.2024.124644

[47] Zeng X., Liao T., Xu L., and Wang Z., "AERMNet: Attention-Enhanced Relational Memory Network for Medical Image Report Generation," *Computer Methods Programs in Biomed*, vol. 244, pp. 107979, 2024. https://doi.org/10.1016/j.cmpb.2023.107979

[48] Zhang J., Shen X., Wan S., Goudos S., and et al., "A Novel Deep Learning Model for Medical Report Generation by Inter-Intra Information Calibration," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 10, pp. 5110-5121, 2023. https://doi.org/10.1109/JBHI.2023.3236661

[49] Zhang K., Zhou R., Adhikarla E., Yan Z., and et al., "A Generalist Vision-Language Foundation Model for Diverse Biomedical Tasks," *Nature Medicine*, vol. 30, no. 11, pp. 3129-3141, 2024. https://www.nature.com/articles/s41591-024-03185-2

[50] Zhang Y., Liu M., Zhang L., Wang L., and et al., "Comparison of Chest Radiograph Captions Based on Natural Language Processing vs Completed by Radiologists," *JAMA Network Open*, vol. 6, no. 2, pp. 2255113, 2023. doi:10.1001/jamanetworkopen.2022.55113

[51] Zhang Z., Wang B., Liang W., Li Y., and et al., "Sam-Guided Enhanced Fine-Grained Encoding with Mixed Semantic Learning for Medical Image Captioning," *in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Seoul, pp. 1731-1735, 2024. https://doi.org/10.1109/ICASSP48485.2024.10446878

[52] Zhao G., Zhao Z., Gong W., and Li F., "Radiology Report Generation with Medical Knowledge and Multilevel Image-Report Alignment: A New Method and its Verification," *Artificial Intelligence in Medicine*, vol. 146, pp. 102714, 2023. https://doi.org/10.1016/j.artmed.2023.102714

[53] Zheng E. and Yu Q., "Evidential Interactive Learning for Medical Image Captioning," *in Proceedings of the International Conference on Machine Learning*, Honolulu, pp. 42478-42491, 2023. https://dl.acm.org/doi/10.5555/3618408.3620195

V S RATNA KUMARI A is working as an Assistant Professor, School of Computer Science and Engineering, VIT AP, Vijayawada, India, she is a proficient educator and keen learner. She received her MCA, Andhra University, Visakhapatnam, India. She was 15 years in different academic layers of qualification. Her areas of research include Expertise in Design and Analysis of Algorithms and Distributed Databases.

Dr. Lalitha Kumari Pappala is working as an Assistant Professor Sr. Gd-1 in the School of Computer Science and Engineering, VIT-AP University has about 15 years of teaching experience. She received her B.Tech degree in Computer Science and Engineering with distinction and M.Tech degree in CSE with distinction from JNTUK, Andhra Pradesh, India. She received Ph.D. degree in NIT Warangal, Telangana state, India. She has published 18 research papers in refereed international journals and conferences. She has received several best paper awards for her research papers at various international conferences. Her areas of research include Machine Learning, Deep Learning, and Image Processing.