Journal of Autonomous Intelligence (2024) Volume 7 Issue 5
doi: 10.32629/jai.v7i5.1542

ORIGINAL RESEARCH ARTICLE

IoT-enabled image captioning with deep learning for healthcare domain

P. Steffy Sherly*, **P. Velvizhy**

Department of Computer Science and Engineering, College of Engineering Guindy, Anna University, Chennai 600025, Tamil Nadu, India

*** Corresponding author:** P. Steffy Sherly, sherlypaul97@gmail.com

ABSTRACT

The ever-increasing volume of medical images greatly strains clinicians who are in the process of reviewing it and writing reports. It would be more efficient and cost-effective if an image captioning model could automatically create report drafts from matching photos, thereby relieving physicians from this tedious work. The Internet of things (IoT) has switched its emphasis from its initial binary concept to that of the Internet of multimedia things (IoMT) because of the explosive rise of multilingual-on-demand data in various sound, footage, picture forms. This work proposed a deep learning-based image caption network (DL-ICN) for healthcare domain. The work originality is shown using DL to identify various class labels of the patient X-ray and ECG images. With the help of bilateral encoder representations from transformers (BERT) method for captioning pictures, a detailed written summary of a person's medical picture may be generated automatically. Results of simulations showed that the proposed model achieved good compression performance, good quality reconstruction and good classification results for image captioning.

Keywords: Bidirectional Encoder Representations from Transformers (BERT); Deep Learning (DL); Image Caption Network (ICN); Internet of things (IoT); Self-Control Differential Evolution (SCDE)

ARTICLE INFO

Received: 12 January 2024
Accepted: 18 February 2024
Available online: 21 March 2024

1. Introduction

Nowadays Hospitals generate massive volumes of medical images from various modalities due to the rapid development of digital health technology based on IoT[1]. One of ambient assisted living (AAL) foundations is remote patient monitoring. In this case, data from various sensors is collected and analyzed to conclude a person's health[2]. Innovative IoT (IIoT) based technologies and services are needed to improve the quality of life for older people[3]. The patient's health may be better understood and predicted if all the medical data obtained from these devices is properly pooled and analysed[4]. It is shown how remote healthcare software using a deep learning classifier may identify anomalies in electrocardiogram (ECG) patterns. Many disorders may be quickly diagnosed and screened for with medical images. They contain useful information regarding several diseases and might be used to spot anomalies. Manual information extraction and description can be time-consuming and laborious because of the complexity and variety of alternative interpretations that may be attributed to a single medical picture[5]. As a result, it is difficult and expensive to process all hospital-generated images promptly. This has a major effect on the team's ability to meet deadlines for submitting reports and ensure they are correct. One option is automated captioning of images, a computer-generated textual description of an image's subject matter written in everyday language[6]. Computer vision (for analysing images) and natural language processing (for writing captions) come together in automatic image captioning[7]. Image captioning is used for several purposes including automated image annotation and labelling, video transcription, security detection, and medical image interpretation[8]. This helps professionals in their everyday jobs and is crucial in computer-aided diagnostic systems, decision-making, and treatment of diseases.

However, captioning medical images is not a simple task. There are several problems with automatically produced reports, including typos and inadequate language explanations. In addition, these reports must conform to stringent clinical standards, which require much knowledge and experience. Such reports, in general, need to adhere to predetermined formats, use specialized medical terminology, and emphasize therapeutically relevant information by providing visual proof instead of just describing the things shown. Notably, traditional captioning algorithms still require refinement to be clinically appropriate since they have trouble generating correct descriptions of medical pictures. Automatically creating text to describe the contents of a picture is the primary focus of the topic of study known as image captioning. In this field, researchers combine techniques from computer vision with natural language processing[9]. Image captioning has several uses, including helping the visually handicapped, assisting with image search, and facilitating interactions between humans and robots. Retrieval-based, template-based, and deep learning-based approaches are some techniques developed for picture captioning[10]. Many studies have lately used deep learning models to categorize individuals with certain conditions[11].

Encoder-decoder architectures with attention mechanisms have been employed in most deep-learning research efforts. A new model where the encoder is a Convolutional Neural Network (CNN) and the decoder is either a Long Short-Term Memory (LSTM) or a transformer. As a result, there is a lot of enthusiasm for applying DL models to the problem of medical picture captioning, which necessitates development of novel methodologies. This has the potential to aid in the rapid exploitation of medical material, prompt delivery of more precise interpretations of results, essential support provided to physicians by reducing their workloads and speeding up clinical processes. Most importantly, this study contributes: DL-ICN was developed for clinical captions to enhance the quality and precision of medical images.

i) The proposed DL-ICN used the BERT approach for image captioning to provide a comprehensive textual analysis of an individual's clinical picture.

ii) It has been shown numerically that the proposed DL-ICN outperforms competing approaches in terms of accuracy and throughput.

The remainder of the article is organized as follows: section 2, covers the literature review; section 3, details the methodology used; section 4, deposit the findings and discussion; and section 5, provides the conclusion.

2. Literature review

In medical image segmentation network model using Atrous Multi-Scale (AMS) convolution, named AMSUnet[12]. AMSE reimagines the reduction encoder that takes advantage of the AMS neural concentration block's atrous and multiple habitats inversion. To enhance feature fusion, design a residual attention mechanism module (i.e., RSC) and apply it to the skip connection. Compared with existing models, proposed model only needs 2.62 M parameters to achieve the purpose of lightweight. Kvasir-SEG dataset, which has a large amount of data and stable experimental results, is selected as the preferred dataset for ablation experiment, and experiments are conducted on two innovative modules, AMSE and RSC, to verify their necessity and excellence. Experimental findings across many datasets show that the developed model performs better in segmenting targets of all sizes.

Simple and effective explainable artificial intelligence (XAI) technique for image text[13]. Deep-learning-based methods address the complexity and difficulties in picture captioning, results have been relatively positive. Azure cognitive service and open-source image captioning model to get image caption. Also, implement XAI image captioning (image to text) using Shapley additive explanations (SHAP). Applies cosine similarity by spaCy and term frequency & inverse document frequency (TF-IDF) to evaluate the sentence similarity. Proposed research 9·ork found that azure cognitive services provides better descriptions for images compared to the open-source image captioning model.

An explainable module for medical image captioning that provides a sound interpretation of attention-based encoder-decoder model by explaining the correspondence between visual features and semantic features[14]. Encoder-decoder models, which consist of two parts working together to create new captions for pictures, are widely used in deep learning-based captioning. Exploit for that, self-attention to compute word importance of semantic features and visual attention to compute relevant regions of the image that correspond to each generated word of the caption in addition to visualization of visual features extracted at each layer of the CNN encoder. Proposed medical image captioning model evaluation, ImageCLEFmed 2021 dataset, which includes three sets: training set composed of 2756 medical images; validation set and test set consisting of 500 and 444 radiology images, respectively. Also, calculate BiLingual Evaluation Understudy (BLEU score) using Python NLTK package. BLEU some visualizations of correctly and wrongly generated captions for the ImageCLEF dataset.

Build an optimized model for histopathological captions of stomach adenocarcinoma endoscopic biopsy specimens[15]. For the image feature extraction subsystem, two evaluations; first, tested 5 different vision models (VGG, ResNet, PVT, SWIN-Large, ConvNEXT-Large) using (LSTM, RNN, bidirectional-RNN) and then compare vision models with (LSTM-without augmentation, LSTM-with augmentation, BioLinkBERT-Large as an embedding layer-with augmentation) to find accurate one. Second, tested 3 different concatenations pairs of vision models (SWIN-Large, PVT_v2_b5, ConvNEXT-Large) to extracted feature vector of the image. For caption generation lingual subsystem, tested a pre-trained language embedding model which is BioLinkBERT-Large compared to LSTM in both evaluations, to select from them most accurate model. Dataset used 34.000 images for training and 5700 images for testing. The performance evaluation metrics are BLEU score and FLOPS. The experiments showed that the best results were obtained when a captioning system was built using the Conv NEXT-Large and PVT_v2_b5 models as an image feature extractor and the Bio-Link BERT-Large language embedding model.

Automatically creating illustrative phrases to accompany a picture is called image captioning[16]. Four different types of learning models were used in this research: First, an image segmentation-based binary classifier called a discriminator; second, an autoencoder; third, a various classes extractor that uses characteristics from both the discriminator and autoencoder to produce keyword labels; and third, a neural network that learns to pair these phrase values with natural language descriptions of skin imaging pathologies. Four, Siamese network learning the textual similarity matching between colloquial description sentences of skin imaging pathology and keywords produced from the multi-class classifier. The experimental results show that the proposed method yields a highest accuracy for the testing data in terms of colloquial language of skin images. The proposed method can significantly relieve the shortage of training personnel and assist hospitals that lack resources for conducting case studies. The results are expected to be feasible and can be applied in actual clinical teaching. The training and testing datasets were collected from DermNet. The performance evaluation metrics are accuracy, Jaccard index (JI), DICE (DSC), sensitivity, specificity, mean squared error (MSE) and mean absolute error (MAE). For medical education in dermatology, findings of this study contribute to the practical value of quantitative indicators and assessments for learning outcomes of medical students.

The autonomous creation of medical imaging reports is where deep learning has recently shown considerable promise [17]. This research looks at current trends and future directions for creating medical imaging reports using deep learning. The dataset, architecture, application, and evaluation of deep learning-based medical imaging report production are all discussed in depth throughout this work. Focus on the deep learning architectures used for generating the diagnostic reports.

3. Proposed methodology

Wearable health monitoring devices based on IoT can track and record the whereabouts and vitals of confined persons in real-time. The terminal monitor can show the current health status of several patients in real-time and alert doctors to potential problems. Research on natural picture captioning has not focused much on the challenge of captioning medical images. Most existing image captioning techniques rely on contextual information from the image itself to create descriptions. However, this cannot be done with medical images because of the need to provide detailed, clinician-style descriptions of the images' contents. Motivated by this, this article suggests generating new captions by exploiting medical notions already connected with photos based on their visual characteristics. The components of the proposed re-trainable network include an EBRT CNN, a long short-term memory (LSTM) model that outputs text and a visual feature encoder that employs a multi-label classifier to classify medical concepts in images. Self-control differential evolution (SCDE) brings about the evolutionary process.

The suggested system has three distinct layers. The IoT devices and the users/patients are separated on the foundational layer. The multi-access edge computing (MEC) node sits smack in the middle, while the cloud's nerve center occupies the opposite extreme. A trustworthy third party (TTP), or entirely reliable outside source, is stationed in the cloud data center. The TTP is a hub for user registration and supplies devices and people with the necessary security measures and access regulations. The edge sends requests to the MEC node, which sends them to the cloud data center. The necessary security information is sent to the MEC node and then provided to the requestors when the cloud service provider has been verified. Specifically, this procedure may be broken down into the six phases listed below (**Figure 1**).

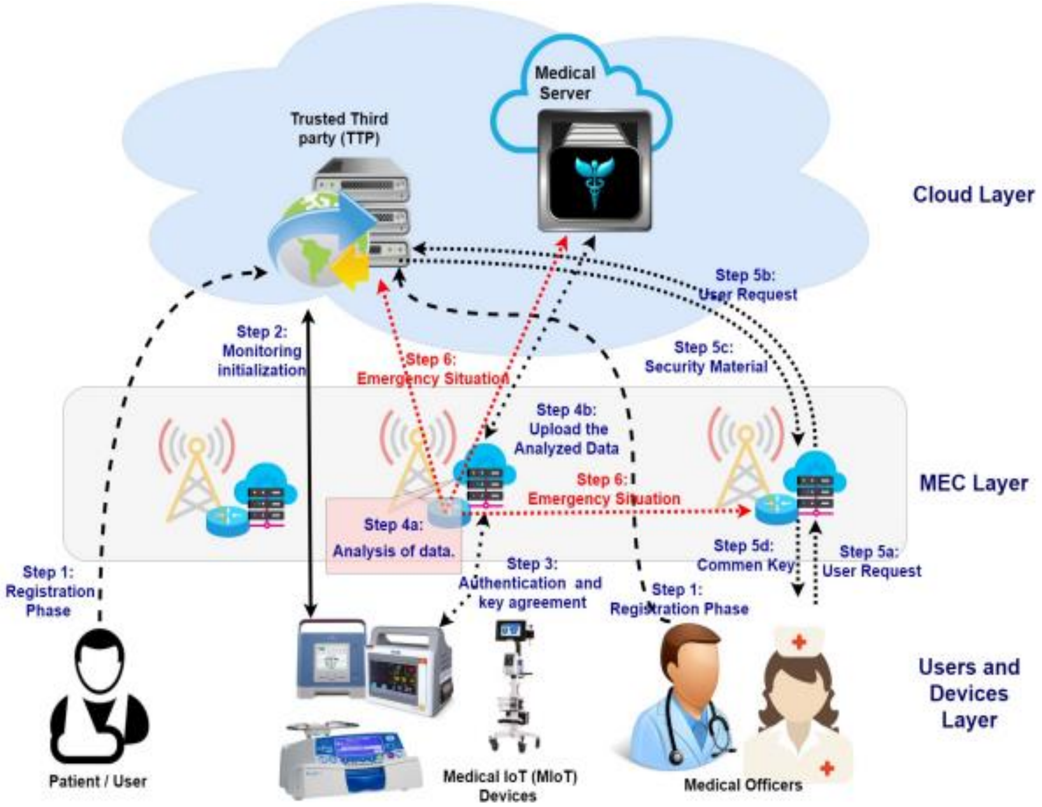
Medical Server
Trusted Third party (TTP)
Cloud Layer
Step 5b: User Request
Step 2: Monitoring initialization
Step 6: Emergency Situation
Step 5c: Security Material
Step 4b: Upload the Analyzed Data
MEC Layer
Step 6: Emergency Situation
Step 4a: Analysis of data.
Step 3: Authentication and key agreement
Step 5d: Commen Key
Step 5a: User Request
Step 1: Registration Phase
Step 1: Registration Phase
Patient / User
Medical IoT (MIoT) Devices
Medical Officers
Users and Devices Layer

Figure 1. Remote monitoring of patients in 5G: Six practical actions.

- Step 1: Registration: In this step, physicians and patients create profiles on the TTP. These profiles will include information such as the doctor's specialties, affiliated hospitals, and patient information such as allergies and blood types. This means that both the user's end (on their phone or smartcard) and the TTP's end have copies of the user's identity, private key, public key, and certificate.
- Step 2: Checking in during the start-up phase: A new analysis profile must be created with associated patient identifier of the analysis (ID_a) and timestamp (Ta) before monitoring with one or more IoT devices may begin for a patient with identification ID_p. Two random parameters, representing the device's dynamic identification DID_i and secret shared key DK_i, are sent between the patient and each IoT device that will participate in the study. The patient sets these two settings and logs their installation time, T_i, under the ID_a of ID_p analysis in the TTP. Additionally, the patient may fill in access characteristics to assist the access control mechanism for the whole analysis profile or particular IoT devices.
- Step 3: Authentication: The IoT and MEC nodes must agree on an authentication method and a shared key. IoT and MEC nodes must first authenticate each other and agree on a new session key before sending encrypted data over the public wireless channel. Therefore, the IoT device requests the MEC node, which transfers it to the cloud hub for verification. In such cases, the cloud data center will relay the necessary protective data to the MEC node. After that, the deduced shared hidden code will allow you to talk to the IoT gadget securely. It is important to note that the cloud center includes the analysis identification ID_a in its response to the MEC node so that the MEC node may aggregate data from several devices that belong to the same analysis/patient.
- Step 4: Information evaluation: The MEC node already knows the identification code of the analysis to which it must add incoming data, thanks to the shared key established between the devices and the MEC node in the previous step. Therefore, ID_a devices may begin sending data to the MEC node in a safe manner so that it can be processed, filtered, collected, aggregated, and interpreted.
- Step 5: Request from the user: The study results may be retrieved by any user, whether patients or doctors, so long as they have the proper access attributes. To do so, the user must first contact the MEC node, which will then transfer the request to the cloud data center. After the cloud service provider verifies the request, it retransmits the necessary security data to the MEC node so that the node and the user may generate a shared secret.
- Step 6: Crisis: If the MEC node's analysis reveals an urgent situation, it will send an alert to the cloud data center, which will then transfer the necessary security resources to the MEC node so it may create a shared key with all relevant parties.

The user uses public key material for the registration, as mentioned above in step 2, and the IoT device holds key material to permit only symmetric key-based activities, as described above in step 1. This is because public-key cryptography is more resource-intensive than symmetric-key techniques in processing and communication. This change is less noticeable since the user can utilize a more capable smartphone or tablet computer. However, minimising security-related expenditures for small IoT devices, like most medical sensors, is important.

This work proposes a deep convolutional network (DCNet) ensemble. The proposed system is created using the VGG16, ResNet152V2, and DenseNet201 models. The Evolving DCNet (EDC-Net) is shown in (**Figure 2**), together with a gated recurrent unit. Over-fitting may be avoided by model assemblage, leading to improved outcomes. It also enhances the performance of controlled models and the effectiveness of identifying features. The input dense layer has 128 neurons. VGG16, ResNet152V2, and DenseNet201 are models used to glean candidate features. These frameworks were initially developed with a large batch of 8 throughout 20 epochs. Researchers have used fully linked layers of size 128 neurons to address memory issues with competitive models with dropouts of 0.2 and 0.25. The learning rate of 0.001 has been used.

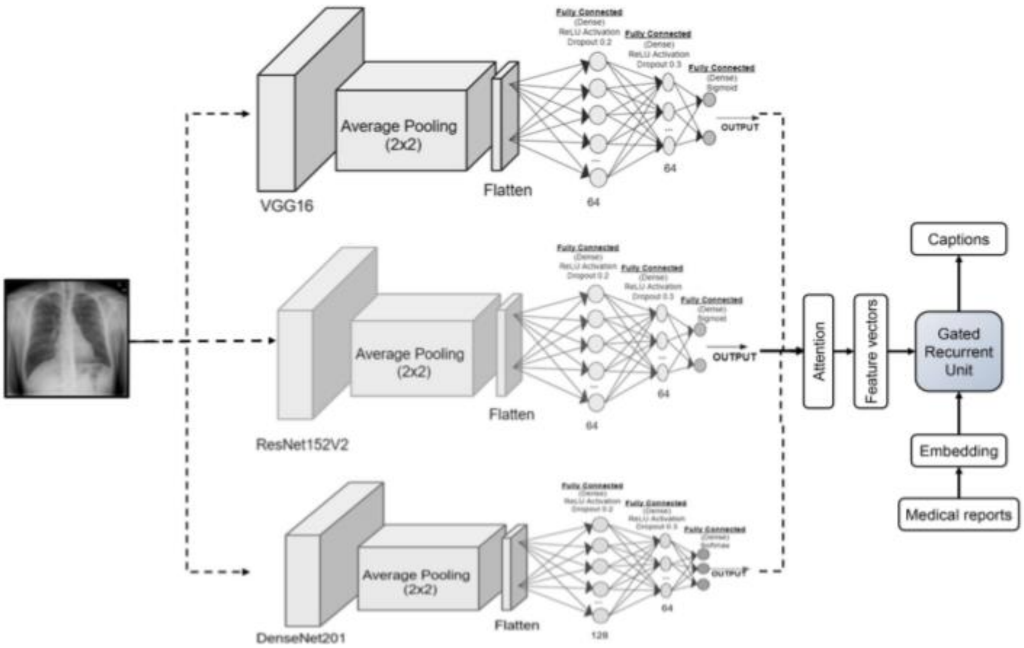
VGG16
Average Pooling (2x2)
Flatten
Fully Connected (Dense) ReLU Activation Dropout 0.2
64
Fully Connected (Dense) ReLU Activation Dropout 0.3
64
Fully Connected (Dense) Sigmoid
OUTPUT
ResNet152V2
Average Pooling (2x2)
Flatten
Fully Connected (Dense) ReLU Activation Dropout 0.2
64
Fully Connected (Dense) ReLU Activation Dropout 0.3
64
Fully Connected (Dense) Sigmoid
OUTPUT
DenseNet201
Average Pooling (2x2)
Flatten
Fully Connected (Dense) ReLU Activation Dropout 0.2
128
Fully Connected (Dense) ReLU Activation Dropout 0.3
64
Fully Connected (Dense) Softmax
OUTPUT
Attention
Feature vectors
Gated Recurrent Unit
Captions
Embedding
Medical reports

Figure 2. A gated recurrent component in a deep transmit group.

a. Connected recurrent unit with gates:

Each recurrent unit may automatically gain multi-scale connections with the help of gated recurrent unit (GRU). GRU employs gating units, similar to LSTM's storage units, to control the flow of information inside the unit. At every given time u, the probability of GRU activation (α_x^a) is an interpolation between the probabilities of applicant activity ($\tilde{\alpha}_x^a$) and subsequent activation (α_{u-1}^a). The Equation (1) for α_x^a is as follows:

$$\alpha_x^a = (1 - \beta_x^a)\alpha_{u-1}^a + \beta_x^a \tilde{\alpha}_x^a \tag{1}$$

here, the activation is monitored and controlled by a status report gateway β_x^a. One way to assess a status report gates is given in Equation (2):

$$\beta_x^a = \sigma(T_v \gamma_x + E_v \alpha_{u-1})^a \tag{2}$$

The reinforced matrix is denoted here by $T_s \gamma_x$. The degree of exposure to which GRU's state is subjected is beyond its control. However, the whole state may be revealed at each cycle. A potential activation $\tilde{\alpha}_x^a$ may be calculated in the following Equation (3):

$$\tilde{\alpha}_x^a = \tan g\big(T_s \gamma_x + E(s_x \odot \alpha_{u-1})\big)^a \tag{3}$$

$\odot$ displays a multiplication by elements here. There is a series of reset gates in s_x. As s_x^a decreases toward 0; by ignoring the previous evaluated state, the reset switch may make the device behave as if it were utilizing the first signal in the supplied series. By comparing the update gate to the reset gate s_x^a, we may calculate in the following Equation (4):

$$s_x^a = \sigma(T_s \gamma_x + E_s \alpha_{u-1})^a \tag{4}$$

The proposed DCNet has a problem with the tuning of its hyper-parameters. The suggested model is evolved in this work using a differential evolution version. The BLUE Multilingual Assessment Understudy serves as the criterion for evaluation. Differential evolution (DE) is a popular method for solving optimization problems because of its many benefits, including high resilience, high performance, and a straightforward structure.

The success of DE relies heavily on the experimental vector generation approach (crossover or mutation) and selection of regulating variables (crossover rate VS, relative magnitude K, and quantity of people IJ). These settings should be chosen for optimal optimization outcomes based on the kind of issue at hand. Parameter setup is a difficult process for any issue. This problem is addressed using DE (SCDE) with self-adaptive parameter control. The theory holds that good qualities should be handed down from generation to generation, while undesirable ones should pick up lessons from the best. Both the solution populations and the variable populations are used in SCDE. Parameters may be used to fine-tune every solution. Population parameters also change over time. SCDE is a hybrid technique that utilizes traditional DE and self-adaptive variable regulation.

Let's assume that DCNet's basic component community is denoted as $V^0 = \{V_1^0, V_2^0, \ldots, V_{IJ}^0\}$ where $V_n^0 = \{K_{n,1}^0, VS_{n,2}^0\}$ and IJ is the population size. The equation denotes the initial population of solutions $R^0 = R_1^0, R_2^0, \ldots, R_{IJ}^0$ where $R_n^0 = r_{n,1}^0, r_{n,2}^0, \ldots, r_{n,A}^0$ where A is the number of independent variables. There will be L_x total generations. Parameters followed the same pattern of evolution as solutions do in DE. At first, a continuously randomized population of parameters is produced between the range [0, 0] and [1, 1]. This is followed by the generation of a mutation parameter $CV_n^{L_x}$ for each $V_n^{L_x}$ through a mutation operator like in Equations (5) and (6):

$$CV_n^{L_x} = V_{s1}^{L_x} + VK\left(V_{s2}^{L_x} - V_{s3}^{L_x}\right) \tag{5}$$

$$CV_n^{L_x} = V_{s1}^{L_x} + VK\left(RV_a^{L_x} - V_{s2}^{L_x}\right) \tag{6}$$

In this case, choose V_{s1}, V_{s2}, and V_{s3} at random from the pool of available parameters. $RV_a^{L_x}$ represents a randomly chosen, beneficial parameter. Then, a crossover operator is used to produce a tail parameter $WV_n^{L_x}$ in Equation (7):

$$WV_n^{L_x} = \begin{cases} CV_{n,m}^{L_x}, \text{if } \left(rand_{n,m} \leq VVS \text{ or } m == m_{rand}\right) \\ V_{n,m}^{L_x}, \text{Otherwise} \end{cases} \tag{7}$$

where n is an integer between 1 and IJ and m is an even number between 1 and 2. The range of uniform random numbers is [0, 1] and is denoted by $rand_{n,m}$. $IVS \in [0, 1]$ represents the current ratio. Ultimately, it uses the selection operator to choose the excellent benchmark for the future. In SCDE, a good parameter of individual $V_n^{L_x}$ is one that aids the $RV_n^{L_x}$ in creating superior progeny $WR_n^{L_x}$. If not, then $V_n^{L_x}$ is a poor choice for the control parameter. If n is a good parameter, then $V_n^{L_x}$ is the selection operator for a good parameter in Equations (8) and (9).

$$V_n^{L_x+1} = \begin{cases} V_n^{L_x}, \text{if } rand(0,1) < \lambda_1 \\ WV_n^{L_x+1}, \text{Otherwise} \end{cases} \tag{8}$$

Else

$$V_n^{L_x+1} = \begin{cases} WV_n^{L_x+1}, \text{if } rand(0,1) < \lambda_2 \\ V_n^{L_x}, \text{Otherwise} \end{cases} \tag{9}$$

where λ_1 and λ_2 choose which parameters to try out with their new values and which to maintain working with their old ones. Algorithm 1 demonstrates the core principles by which SCDE operates. Line 1 initially generates a population of solutions, whereas line 2 generates a population of parameters. The generation number L_x is set to 1 on line 3. The symbol denotes the function evaluation count K_{eval}. When $K_{\max}$ approaches $K_{\max}$, as shown on line 5, the algorithm terminates. LH remembers which people made up the excellent parameter (line 6). The value is set to 0 by default. From line 7 through line 16, a population of solutions is evolved. In line 7, apply the mutation operator to an individual $V_n^{L_x}$ as a parameter to produce a mutant vector $RC_n^{L_x}$. The Equation (10) is represented as following:

$$RC_n^{L_x} = R_{s1}^{L_x} + K_{n,1}^{L_x}\left(R_{s2}^{L_x} - R_{s3}^{L_x}\right) \tag{10}$$

where R_{s1}, R_{s2}, and R_{s3} are picked at will from the population of the solution. The notation denotes the factor of magnification $K_{n,1}^{L_x} \in V_n^{L_x}$. The $V_n L_x$ trial vector $WR_n^{L_x}$ is obtained by applying the crossover operator in line 8. The Equation (11) is represented as following:

$$WR_n^{L_x} = \begin{cases} RC_{n,m}^{L_x}, \text{if } (rand_{n,m} \leq VS_{n,2} \text{ or } m == m_{rand}) \\ R_{n,m}^{L_x}, \text{Otherwise} \end{cases} \tag{11}$$

where $n = [1, 2, \ldots, IJ]$ and $VS_{n,2} \in V_n^{L_x}$. The optimal answer is chosen using a selection operator (lines 10–15). The associated $V_n^{L_x}$ is a marked selection of regulating variables with a 1 (i.e., $val = 1$) (line 9) if the fitness of $WR_n^{L_x}$ is higher than that of $R_n^{L_x}$. Line 10 of the LH also has the addition. The parameter $V_n^{L_x}$ is marked as invalid if it is less than 1 (line 17). The next step is to use mutation, crossover, and selection operations to develop the parameter population (lines 19–36). To test out the new values, improper initializations of the parameters (line 31) are used if $LH = 0$.

Algorithm 1 Self-control differential evolution ensemble model

```
1: Create a prototype of the solution ($R^0$) and distribution of parameters ($V^0$)
2: Set $L_x = 1$ and $K_{max} = 0$
3: While $K_{eval} < K_{max}$ do
4: Set $LH = 0$
5: For $n = 0$ to $IJ$ do
6: Vector mutation $RC_n^{L_x}$ is receivedfrom Equation 10 and $V_n^{L_x}$
7: Trial vector $WR_n^{L_x}$ is obtained using Equation 11 and $V_n^{L_x}$
8: If $k(WR_n^{L_x}) \geq k(R_n^{L_x})$ then
9: $R_n^{L_x+1} = WR_n^{L_x}, val(n) = 1$
10: Put $V_n^{L_x}$ into $LH$
11: Else
12: $R_n^{L_x+1} = R_n^{L_x}, val(n) = 0$
13: End
14: End
15: For $n = 1$ to $IJ$ do
16: If $rand(0,1) < \lambda_1$ then
17: $V_n^{L_x+1} = V_n^{L_x}$
18: Else
19: Generate $WR_n^{L_x}$ using Equation (5) and Equation (7)
20: $V_n^{L_x+1} = WV_n^{L_x}$
21: End
22: Else
23: If $rand(0,1) < \lambda_2$ then
24: If $LH \neq 0$ then
25: Generate $WV_n^{L_x}$ using Equations (6) and (7)
26: $V_n^{L_{x+1}} = WV_n^{L_x}$
27: Else
28: Initialize $V_n^{L_{x+1}}$ randomly
29: End
30: Else
31: $V_n^{L_{x+1}} = V_n^{L_x}$
32: End
33: End
34: End
35: $L_x = L_{x+1}$
36: End
```

b. Embedding language:

For linguistic demonstrations, one can put a model of language based on the BERT database (BBLM) that equates to the Equations (12)–(14) below:

$$lf = BERT(W) \tag{12}$$

$$S = MaskedAttention(FF1(lf) + pos) \tag{13}$$

$$W = log(sotfmax(FF2)) \tag{14}$$

In such case $W = (< bos >, W1, W2, \ldots, W_M)$ is input sequences; $< bos >$ is an acronym for initial sentence; this token is appropriate for the “introductory” anticipate the captions first using the decoder; the position encoding of words is $pos \in \mathbb{R}^{d_{bert}}$. Embedded representation of sequences (position encoding). Component-specific position vectors FF1 and FF2 are the first and second frames, respectively, feed-forward networks with hidden nodes, including a ReLU-activated two-layer linear network. These networks rely on feed-forward knowledge of the model based on transformers, which handle focus and presentation than the preceding multi self-attention node. version; $If \in \mathbb{R}^{d_{bert}}$ is the BERT model's output $S \in \mathbb{R}^{d_{bert}}$ is the result of covert processing. W is the log softmax probability in this module of word-prediction distribution.

Considering that VieCap4H is a Vietnamese dataset, we use the HuggingFace-hosted vinai/phobert-base model to pre-train a BERT-based language model. In addition, the phobert-base model is somewhat designed and optimized for training quickly on the tiny VieCap4H dataset, allowing us to run more trials. It also gives BARTPho-syllable and BARTPho-word, as well as PhoBERT-large. Yet, it does not appear to work well with pre-trained models. Possible cause: VieCap4H's very modest sample size makes large-scale systems inappropriate. The linguistic device "attention" portrays a single reference sequence.

c. Methodology of RSTNet:

Specifically, during training, we use the grid augmented (GA) and adaptive attention (AA) modules of the RSTNet model—a suggested Transformer-based architecture—to improve the model's performance. This is the strategy we settled on for testing the VieCap4H. Dataset since it is a novel approach with two novel modules. Grid characteristics are used throughout the design to lower the computational complexity of the architecture. Additionally, RSTNet pre-trains a BERT-based model to get language signals and utilises the grid features and random embedding vectors to train the Transformer-based model. Then, the Adaptive Attention module integrates the encoder's output (visual encoded features), the decoder's (hidden states), and the linguistic signals to predict the following word. It swaps out its BERT-based model with our PhoBERT-base variant so that it may be tailored to each nation.

d. An enhanced grid (GA):

The study significantly altered the transformer's conventional attention mechanism to obtain the relative geometry matrix $\lambda^g \in \mathbb{R}^{N \times N}$ that represents the spatial relationships between grids. To properly integrate the new spatial data into the model's computations, this modification was necessary. Self-attention mechanisms are the mainstay of the typical Transformer architecture, enabling the model to balance the relative relevance of various input components during prediction. However, a basic self-attention mechanism might not be enough for activities involving geographic or grid data. As a result, this work add's a grid-specific attention mechanisms to the Transformer to increase its functionality and handle the intricate interactions between grids.

e. Adaptive attention (AA):

The authors discovered instances where the next word prediction relied more on linguistic context than visual attributes. It is suggested that the AA module use language presentations (from the mask attention module at the BERT model), visual cues from the encoder output, and hidden states to anticipate the next word probabilities rather than relying only on the hidden states of the decodes. In particular, the decoder output at timestep t is fed into another instance of the mask multi-head attention to generate an attention feature, which is subsequently used as a query. A key is any encoder-generated visual signal at the current time step t. At timestep t, the BERT-based language model outputs a linguistic signal that is quantified. The Multihead fits the criteria of the query, key, and value. Focus on trying to guess the following word. An excellent illustration of the inference and training stages (**Figure 3**).

Natural picture captioning methods based on object recognition have inspired the description a captioning strategy for medical image captioning that blends derived visual characteristics with related medical concepts. Indeed, it is not possible to use object detection algorithms on medical pictures since they do not provide accurate findings. However, certain pre-trained algorithms may be used to extract medical ideas linked to medical imagery.

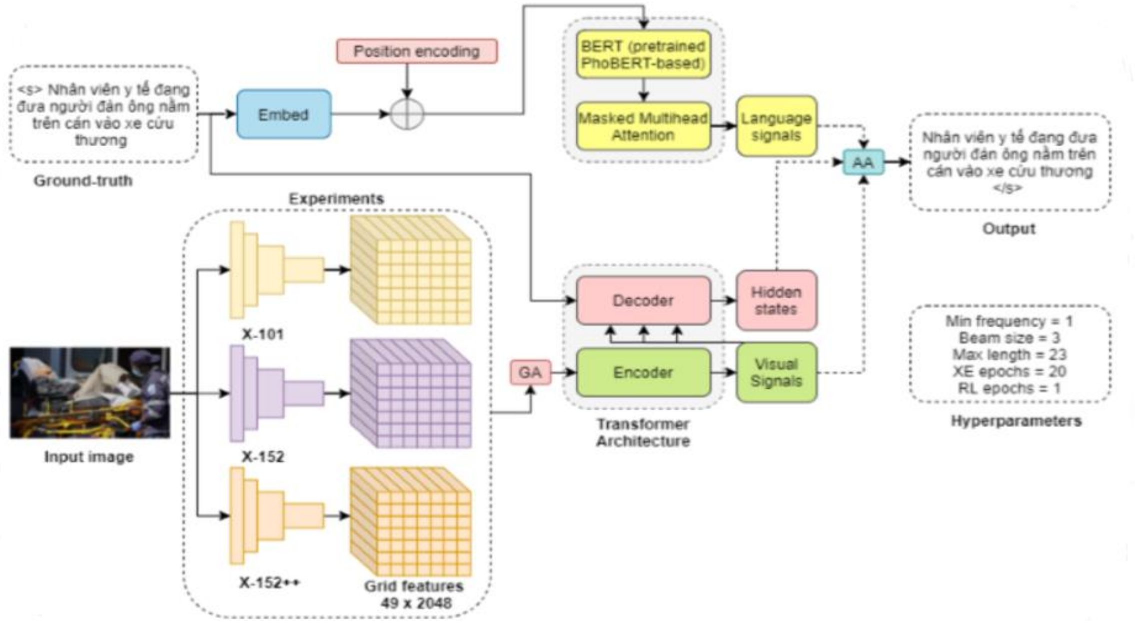

Figure 3. Experimental procedure summary.

The results from such systems have the potential to enhance the accuracy of already available medical image captioning systems. It advocates combining visual and semantic elements to provide alternative captions. Some pre-trained networks are used to derive visual characteristics from photos. Medical concept detection for medical pictures is used to calculate semantic characteristics. CNN networks are used to obtain characteristics from optical and lexical samples, and then a multi-label decoder is used to identify the concepts. In the end, a long short-term memory (LSTM) network is implemented for linguistic output, with beam search used to improve the prediction of suitable phrases for use in the caption. Detailed explanations of the procedures (**Figure 4**).

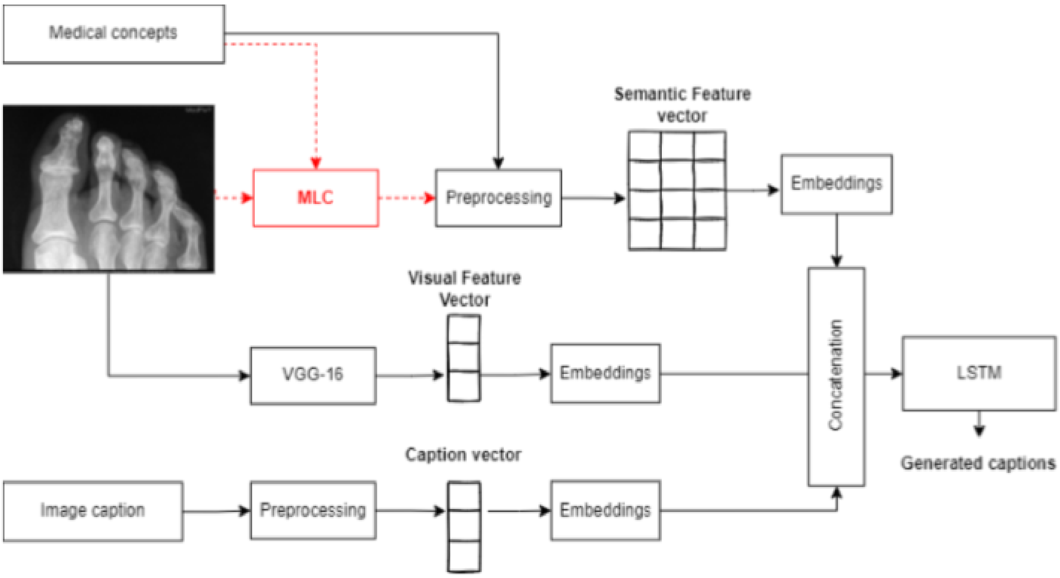

Figure 4. Multi-label LSTM feature vector creation and training.

f. Encoding visual features:

As a first step, it is proposed or use an already-trained CNN model for feature extraction. Because of its modest size, high performance, and extensive training on the massive ImageNet dataset, VGG-16 model is employed for the classification needs. After feeding the medical photos into the model's 16 levels (**Figure 3**), the final classification layer is removed to reveal a 4096-element feature vector. The model learned these properties while attempting to forecast the picture class and differentiate visual content. To fit the images into the VGG-16 model, they were pre-processed. They were scaled and normalized to work in the encoder and then enhanced using conventional methods.

g. Pre-processing of text:

This work pre-processed the captions to remove unnecessary words and tidy up the content. To be more precise, it deleted stop words, filtered out punctuation, and computed word stems after tokenizing each caption and changing all letters to lowercase. Each caption now also includes the words "start" and "end" to denote where the statement begins and ends. The NLTK package is used for this pre-processing. One caption in the training set could be no longer than 50 words; thus, the longer ones were utilized to fill the set. The semantic content of each statement was also captured by calculating embeddings from these captions. Finally, a 50-by-50-by-1 vector was used to represent each picture and encode its caption.

The proposed model combined visual and semantic information into a single joint feature to generate captions. Thus, the medical ideas linked with the photos were processed to get semantic characteristics. Similar processing was used with the captions. Concepts of distinctive designations (CUIs), such as C1306645 for "Plain X-ray", represent the medical ideas made available to the public. Tokenizing, lowercasing, stemming, and removing stop words were all applied to the many terms that make up the UMLS. This led us to investigate ten distinct concepts, each encoded using a vector of size 9 (consisting of nine words) to ensure that the words accurately reflected the visuals. After tokenization, the maximum length of a CUI was 9 words, and 10 CUIs were optimal for each notion in the training set. We padded the ending sequences if the notion needed fewer words. If less than 10 CUIs were connected with a picture, we used the last CUI's vector for the remaining places. The semantic characteristics consisted of embeddings computed from medical concepts.

h. Learning new words:

This language was constructed through a number of phases. The process began with turning a corpus of text into a set of tokens, where each token stood for a unique linguistic unit, which may be a word or even a smaller linguistic piece. These tokens demonstrate the language's adaptability and variety as they were created from captions and conceptual ideas. The construction of a lexicon, where each distinct term or symbol was given a numerical value depending on its particular place within the dictionary, was the next step in this linguistic progression. The purpose of this numerical assignment was probably to give the language a clear structure and order. The positional value may have been influenced by linguistic considerations, word usage patterns, context importance, or other variables. The numerical values assigned to these terms may be used within the created language for a number of functions, including indexing, sorting, or creating a hierarchy of priority.

i. Identifying medical ideas through multi-label analysis:

The proposed model advocates using a multi-label classifier to identify and categorize medical picture ideas. Then, it trained a condensed version of the VGG-16 network's CNN to do this. The medical photos served as inputs to the model, while the various classes that resulted from identifying ideas based on visual attributes served as outputs. For each picture, we chose the 10 most common associations. The anticipated ideas underwent further pre-processing, serving as the building blocks for our semantic feature encoding. This article can explain how the multi-label classifier is set up (**Figure 5**).

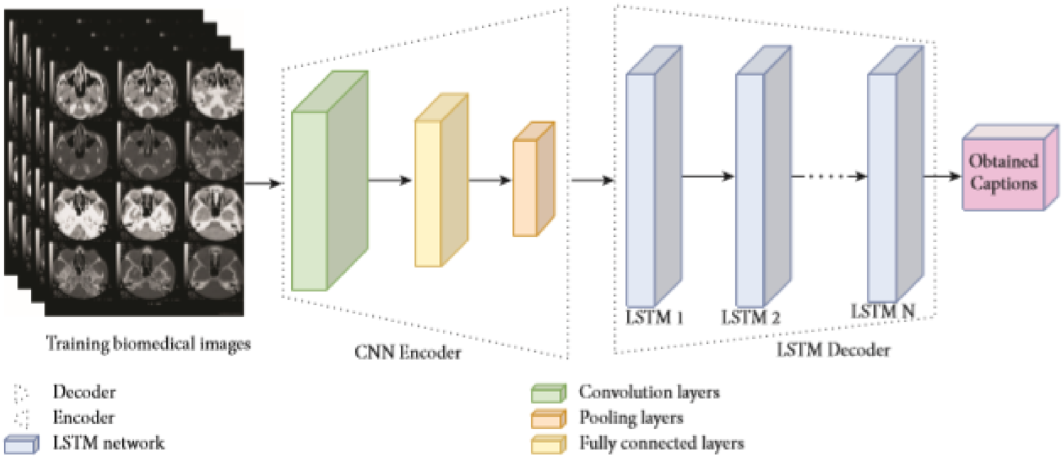

Figure 5. DL-ICN architecture.

The encoder-decoder method is considered to introduce a strategy for visual attention. The decoder can automatically focus on the most important parts of a medical picture to provide a good description. **Figure 5** represents a proposed DL-ICN.

To get G vectors in B dimensions, this model uses a convolutional neural network (CNN) as an encoder. Each vector stands for a mask in the diagnostic picture. The feature vectors are scored using the output of the convolutional layer in Equation (15).

$$s = \{s_1, \ldots, s_G\}, s_n \in \mathbb{R}^B \tag{15}$$

The decoder section makes use of LSTM to provide descriptions. The context vector is calculated in Equation (16),

$$w_t = \sum_{n=1}^{B} \alpha_{tn} c_n \tag{16}$$

where w_t describes the implementation of the attention technique, andα is calculated for each iteration t of the algorithm. The attention weight vector at iteration w_t is expressed by the expression $\alpha_t \in \mathbb{R}^B$. A neural network may be used to approximate $\sum_{n=1}^{B} \alpha_{tn} = 1$. c is the formula for a SoftMax activation function in Equation (17).

$$\alpha_{tn} \propto \exp\{d_{ctt}(c_n, j_{t-1})\} \tag{17}$$

Accordingly, this may characterize the suggested attention encoder-decoder model in the following Equations (18)–(20).

$$c = Encoder(N) \tag{18}$$

$$w_t = \sum_{n=1}^{B} \alpha_{tn} c_n, \alpha_{tn} \in \mathbb{R}, c_n \in \mathbb{R}^G \tag{19}$$

$$\begin{gathered} i_t = Q_r A_t, t \in \{0, \ldots, X-1\}, \\ N_{t+1} = Decoder(i_t, w_t), t \in \{0, \ldots, X-1\} \end{gathered} \tag{20}$$

However, SCDE-based LSTM is quite picky about the very first settings. As a result, SCDE is used to fine-tune the baseline properties of SCDE-based LSTM. The normal distribution is first used to generate a random population sample for further mathematical and other technical information on SCDE and hyper-parameter tuning concerns. Then, the solutions that are not dominated are calculated and included in the Pareto set. The fitness score is then calculated. Then, operators like crossover and selection are employed to provide novel solutions. Once again, we calculate the fitness of the calculated solutions. At last, the solutions that are not dominated are added back to the Pareto set. These procedures will continue indefinitely until and until the termination conditions are met.

4. Numerical results

As a greater number of medical images need to be examined and reported, clinicians are under increasing stress. If a computerized image captioning system could automatically generate report drafts from corresponding images, saving doctors time and money, this would be a significant time and labour savings. As demand for streaming sound, video, and still image content has skyrocketed, the initial scalar vision of the IoT paved the path towards the virtual world of multisensory matters. Therefore, this work employs a DL-ICN to do medical picture captioning. The paper's novel approach is shown by using deep learning (DL) to determine the X-ray and electrocardiogram pictures' respective class labels. The evolutionary process is triggered by self-control differential evolution (SCDE). The BERT method (bidirectional encoding representation using transducers) for the problem of captioning images. An expert textual description of the patient medical picture is supplied. BERT is based on the field of computational linguistics. The simulation results demonstrated that the proposed model successfully achieved high standards of compression performance, high-quality reconstruction, accurate classification, and accurate picture captioning.

4.1. Accuracy analysis

Accuracy comparisons between the proposed SCDE- LSTM based DL-ICN and other medical image captioning models (**Figure 6**). It is inferred that proposed model yields 92.8% higher accuracy when compared with other models.

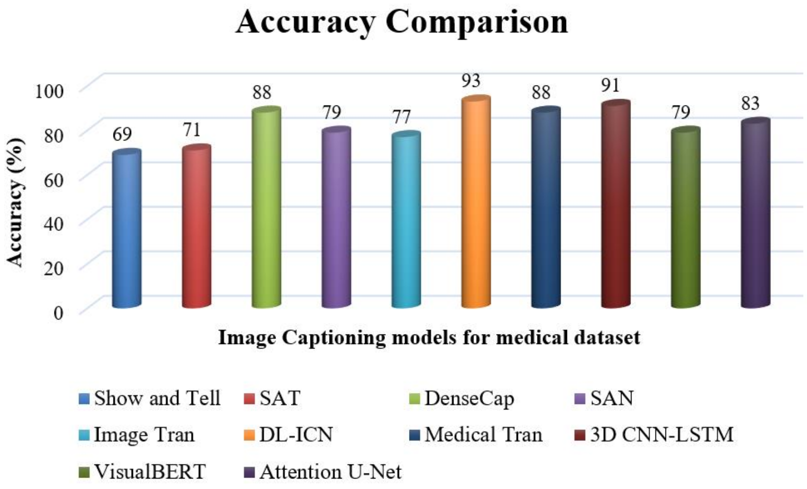

Figure 6. Accuracy comparison.

The accuracy performance may be improved by further training the model and fine-tuning the hyperparameters. Medical ideas and picture captions are good candidates for data augmentation to expand the available text corpus. Caption quality might be improved by exploring the pre-trained embedding model BERT, like those educated on a huge medical database, to include various medical languages in the training data. A specific image's related ideas are then predicted based on the two tiers of information. Moreover, clinically relevant data need to be retained during the pre-processing phase.

4.2. Error analysis

The error analysis of the SCDE-based LSTM (**Figure 7**). The observed root mean square error is 13.1 when the epoch is 7. Particle data from epoch 7 is utilized to fine-tune SVM's tuning parameters. At epoch 7, it was discovered that the root mean square error was zero across all datasets (training, testing, and validation). Thus, the best parameters are used by SCDE-based LSTM to train the medical image captioning model. The SCDE-based LSTM's validation, mean (mu), and gradient tests are shown in **Figure 7**. A gradient is determined to be 10.2, and mu is found to be 9.5 when validation tests are run 6 times. As a result, SCDE-based LSTM can do well regarding caption recognition. Furthermore, *mu* = 9.5 demonstrates that the overfitting issue is absent in SCDE-based LSTM.

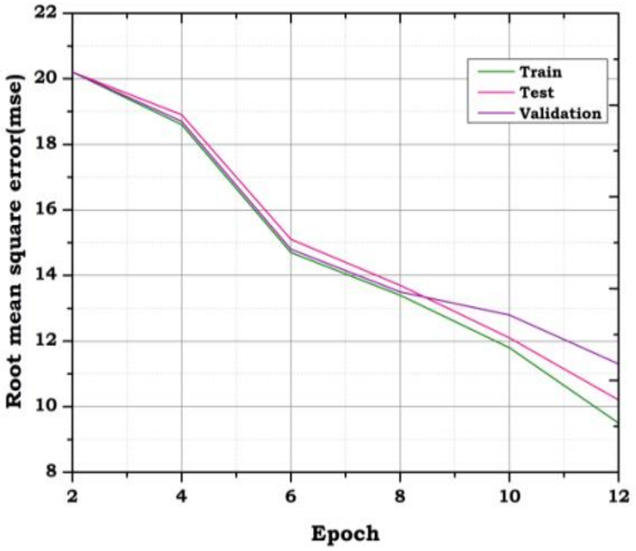

Figure 7. Error analysis.

4.3. Visual caption analysis

Figure 8a displays the accurate transcriptions. It turns out that the assumed subtitles are a dead-ringer for the actual ones. As a result, we got a perfect BLUE rating for captions like these. Because we have treated it as a classification issue, the true projected class is also shown there. It's conclusive evidence that DL-ICN can successfully give captioning for medical photos.

The misread subtitles (**Figure 8b**). Captions that were projected were quite different from those that were used. As a result, it is guaranteed at least a BLUE rating for such captions. The misclassified group is also shown since we've treated it like a classification issue. This demonstrates that DL-ICN does not always produce accurate captions, especially when picture visibility is low. Images have been accurately categorized using the proposed model, and the BERT model has created captions. Moreover, some captions include numerical data or punctuation with significant semantic value. However, the pre-processing step of excluding punctuation, stop words, and certain characters might lead to a loss of information that compromises the new caption's meaning and generation. Excerpts from the original captions that the removal of some numbered tokens has modified (**Figure 8a,b**).

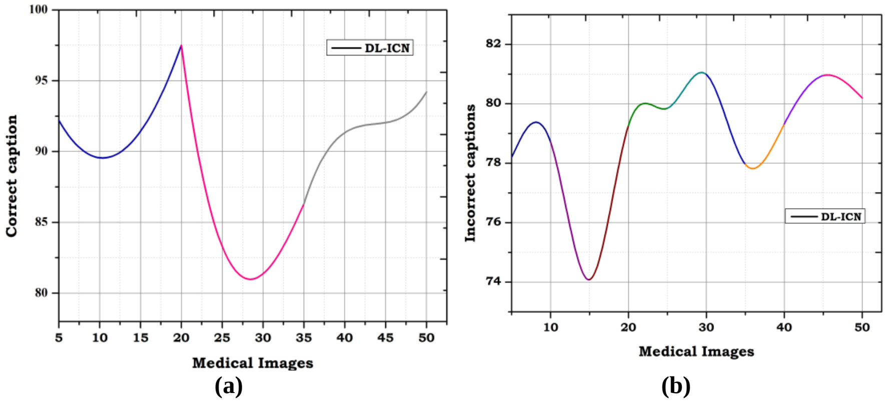

Figure 8. (a) correct caption prediction; **(b)** incorrect caption prediction.

4.4. Analyzing losses during training and testing

This article discusses the use of loss curves in the training and testing data analysis. The loss analysis of VGG16 during training and testing as a function of epoch count (**Figure 9a**). The VGG16 performs best with a training loss of 0.312 and a testing loss of 0.212. This case illustrates the impact of over-fitting.

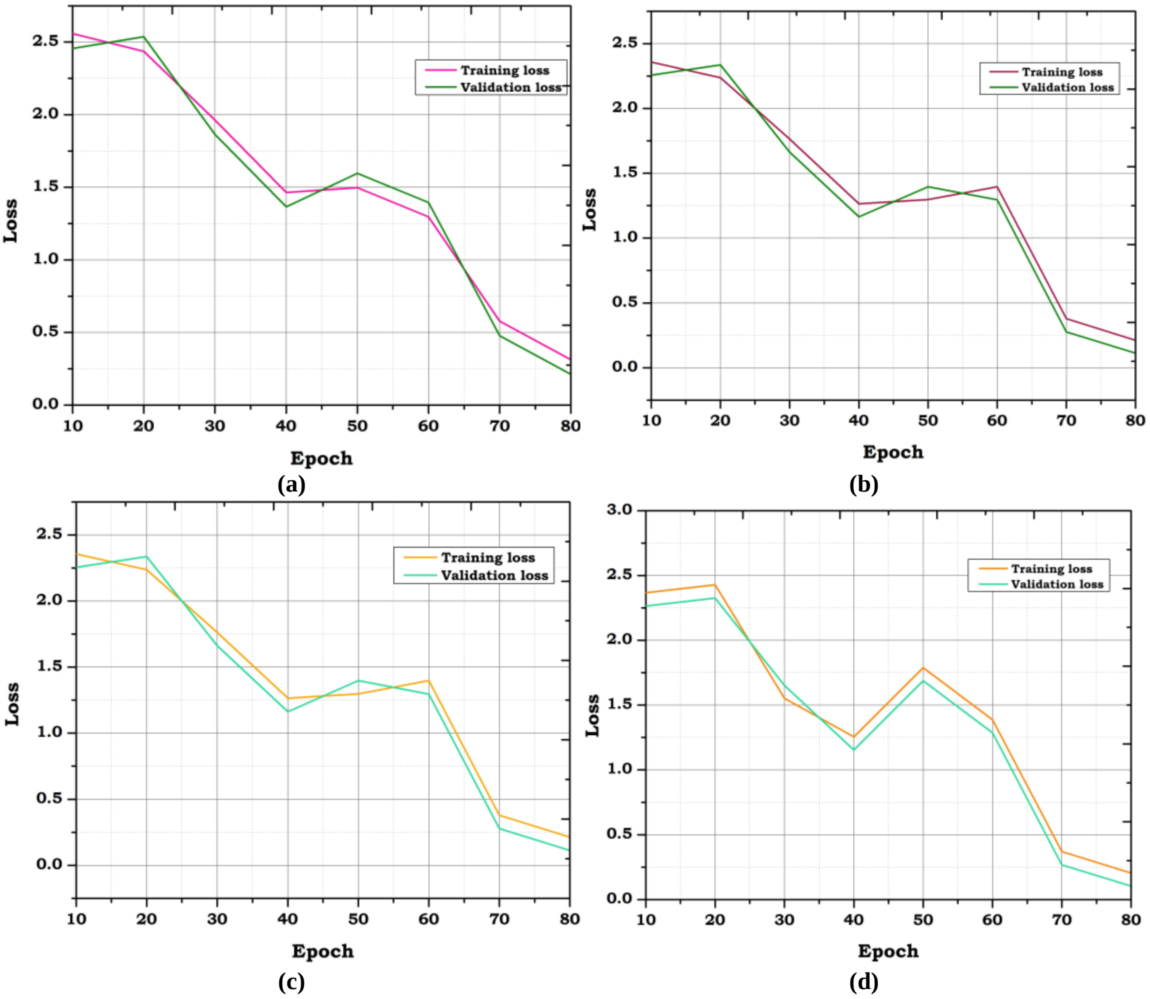

Figure 9. (a) training and testing loss analysis of VGG16; **(b)** training and testing loss analysis of ResNet152V2; **(c)** training and testing loss analysis of DenseNet201; **(d)** training and testing loss analysis of DL-ICN.

The ResNet152V2 loss analysis during training and validation (**Figure 9b**). The best results for training and validation loss are obtained with RESNET152V, with values of 0.212 and 1.112, respectively. Overfitting is shown to have an effect. However, it outperforms VGG-16 in tests. DenseNet201's loss analysis during training and validation (**Figure 9c**). DenseNet201 is the least impacted by the over-fitting issue, with optimal training and testing loss values of 0.214 and 0.113, respectively.

Nonetheless, the convergence curve may be sharpened further. Training and validation loss analyses for DL-ICN (**Figure 9d**). The results show that DL-ICN strives for the smallest possible training loss. Also, the gap between the optimum training and validation loss values, 0.204 and 0.103, is less for DL-ICN, making it more resistant to the over-fitting issue.

4.5. Performance analysis

The bilingual evaluation understudy (BLEU) score is a widely used metric in the field of natural language processing and machine translation to evaluate the quality of machine generated text, such as machine translation or image captions. It was developed to assess the similarity between a reference human-generated sentence and a candidate (machine-generated) sentence. In **Table 1**, BLEU score ranges from 0 to 1, with higher scores indicating better quality and more similarity between the two sentences. BLEU is based on the concept of precision, which measures the proportion of words in the candidate sentence that are also present in the reference sentence. The BLEU score is calculated using the following Equation (21):

$$BP \times exp(1/n \times \sum(log(p\_i))) \tag{21}$$

where brevity penalty (*BP*) is a factor that penalizes short candidate sentences when they don’t match the reference sentences’ length. n is the n-gram order (usually 1, 2, 3, or 4) that determines the precision, and it considers both unigrams, bigrams, trigrams, and so on. The variable *p_i* represents the precision for each *n*-gram. The BLEU score is a valuable tool for comparing and evaluating the performance of machine translation systems and other text generation models, providing a quantitative measure of how closely the generated text aligns with human references.

Table 1. Performance analysis comparison for other image captioning existing models and proposed DL-ICN.

S.no	Model name	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	CIDEr
1	DL-ICN	0.93	0.76	0.62	0.51	0.77	1.33
2	InceptionV3 + LSTM	0.75	0.60	0.50	0.45	0.65	1.20
3	ResNet + GRU	0.78	0.63	.052	0.47	0.68	1.25
4	Transformer	0.80	0.66	0.54	0.49	0.70	1.20
5	Show, attend, tell	0.76	0.62	0.51	0.67	0.67	1.22
6	CNN-LSTM Hybrid	0.73	0.58	0.48	0.63	0.63	1.18
7	VGG16 + GRU	0.77	0.61	0.49	0.66	0.66	1.23
8	BERT for images	0.79	0.64	0.53	0.69	0.69	1.27
9	MobileNetV2 + LSTM	0.74	0.59	0.47	0.64	0.64	1.19
10	Inception-ResNet + GRU	0.76	0.62	0.51	0.67	0.67	1.21

a. Bilingual evaluation understudy (BLEU) analysis

Bilingual evaluation understudy (BLEU) comparisons between the proposed SCDE-LSTM based DL-ICN and other medical image captioning models (**Figure 10**). It is inferred that proposed model yields 0.93% higher BLEU-1 when compared with other models.

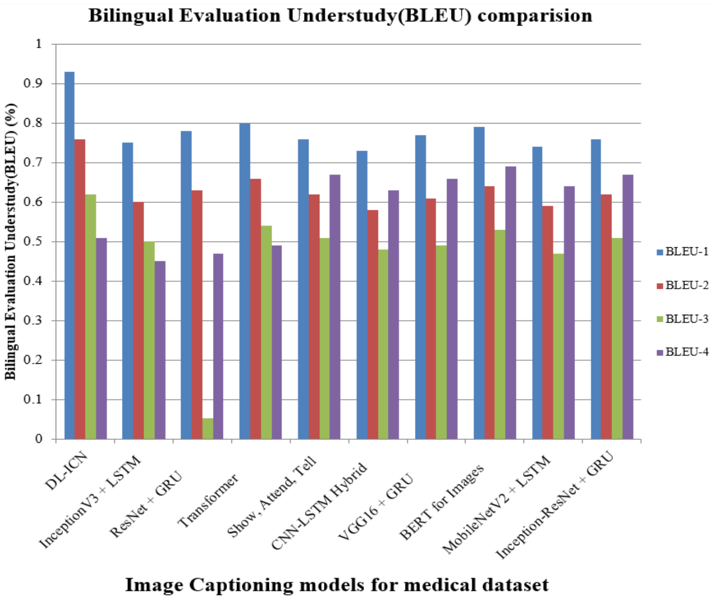

Figure 10. Bilingual evaluation understudy (BLEU) comparison.

The bilingual evaluation understudy (BLEU) comparison performance may be improved by further training the model. Medical ideas and picture captions are good candidates for data augmentation to expand the available text corpus. Caption quality might be improved by exploring the pre-trained embedding model BERT, like those educated on a huge medical database, to include various medical languages in the training data. Moreover, clinically relevant data need to be retained during the pre-processing phase.

b. Metric for evaluation of translation with explicit ORdering (METEOR) analysis

Metric for evaluation of translation with explicit ORdering (METEOR) is a metric for the evaluation of machine translation output. The metric is based on the harmonic mean of unigram precision and recall, with recall weighted higher than precision. METEOR score for this pair of translations is computed as follows. First unigram precision (P) is computed as the ratio of the number of unigrams in the system translation that are mapped to the total number of unigrams in the system translation. Similarly, unigram recall (R) is computed as the ratio of the number of unigrams in the system translation that are mapped to the total number of unigrams in the reference translation. Next compute Fmean by combining the precision and recall via a harmonic-mean that places most of the weight on recall.

The resulting formula used is Equation (22):

$$\text{Fmean} = \frac{10PR}{R + 9P} \tag{22}$$

First, all the unigrams in the system translation that are mapped to unigrams in the reference translation are grouped into the fewest possible number of chunks such that the unigrams in each chunk are in adjacent positions in the system translation, and are also mapped to unigrams that are in adjacent positions in the reference translation. In the other extreme, if there are no bigram or longer matches, there are as many chunks as there are unigram matches. The penalty is then computed through the following Equation (23):

$$\text{Penalty} = 0.5\left(\frac{\text{Number of chunks}}{\text{Number of unigrams matched}}\right)^{3} \tag{23}$$

Finally, the METEOR score for the given alignment is computed as follows Equation (24):

$$\text{Score} = \text{Fmean}1(-\text{Penalty}) \tag{24}$$

METEOR comparisons between the proposed SCDE-LSTM based DL-ICN and other medical image captioning models (**Figure 11**). It is inferred that proposed model yields 0.77% higher when compared with other models.

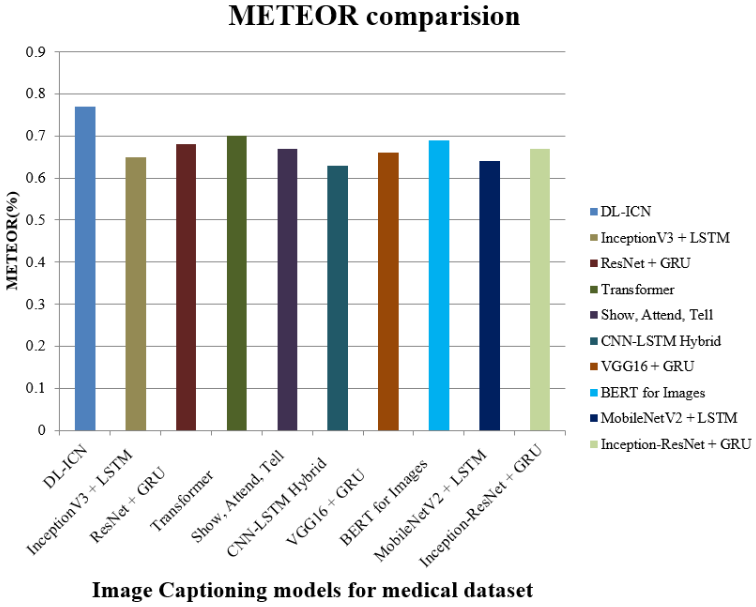

Figure 11. Metric for evaluation of translation with explicit ORdering (METEOR) comparison.

The metric for evaluation of translation with explicit ORdering (METEOR) comparison performance may be improved by further training the model. Medical ideas and picture captions are good candidates for data augmentation to expand the available text corpus. Caption quality might be improved by exploring the pre-trained embedding model BERT, like those educated on a huge medical database, to include various medical languages in the training data. A specific image's related ideas are then predicted based on the two tiers of information.

c. Consensus-based image description evaluation (CIDEr) analysis

The CIDEr metric measures the similarity between a generated caption and the reference captions, and it is based on the concept of consensus: the idea that good captions should not only be similar to the reference captions in terms of word choice and grammar, but also in terms of meaning and content. The CIDEr metric is computed as follows:

(1) First, a set of reference captions is provided for each image. These captions serve as the ground truth for the evaluation.

(2) The generated caption is compared to each reference caption using the BLEU score, which measures the n-gram overlap between the generated caption and the reference captions.

(3) The BLEU scores are then modified using an IDF (inverse document frequency) weighting, which gives more weight to words that are rare in the reference captions but appear in the generated caption.

(4) Finally, the weighted BLEU scores are averaged over all reference captions to produce the final CIDEr score.

CIDEr comparisons between the proposed SCDE-LSTM based DL-ICN and other medical image captioning models (**Figure 12**). It is inferred that proposed model yields 1.33% higher when compared with other models.

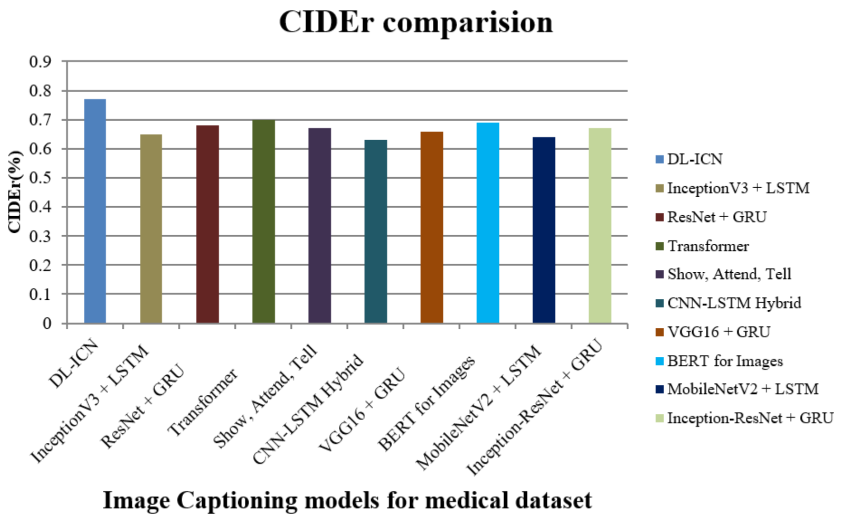

Figure 12. Consensus-based image description evaluation (CIDEr) comparison.

The consensus-based image description evaluation (CIDEr) comparison performance may be improved by further training the model. Medical ideas and picture captions are good candidates for data augmentation to expand the available text corpus. Caption quality might be improved by exploring the pre-trained embedding model BERT, like those educated on a huge medical database, to include various medical languages in the training data.

5. Conclusion

IoT Technology can be worn to monitor the health and whereabouts of restricted individuals in real-time. The terminal monitor can display many patients' current health statuses and send alerts to physicians if any concerns arise. Even though the natural image captioning is considering as facile technique, there is still a lot of uncharted challenge in case of medical image captioning. Several existing image captioning systems now leverage in-image non-visual components to build descriptions but, the need for extensive, clinician-style explanations of the contents of medical photographs prevents this from happening. As a result, this work proposes a method for automatically creating new captions for medical images by mining pre-existing associations between medical concepts and the photos' visual qualities. Using a multi-label classifier, the visual feature encoder, the BERT CNN, and the long short-term memory (LSTM) model that outputs text make up the whole trainable network.

Author contributions

Conceptualization, PSS and PV; methodology, PSS; software, PSS; validation, PSS and PV; formal analysis, PSS; investigation, PSS; resources, PSS; data curation, PSS; writing—original draft preparation, PSS; writing—review and editing, PSS; visualization, PSS; supervision, PSS; project administration, PSS; funding acquisition, PV. All authors have read and agreed to the published version of the manuscript.

Conflict of interest

The authors declare no conflict of interest.

References

1. Singh S, Rathore S, Alfarraj O, et al. A framework for privacy-preservation of IoT healthcare data using Federated Learning and blockchain technology. Future Generation Computer Systems 2022; 129: 380–388. doi: 10.1016/j.future.2021.11.028
2. Mehbodniya A, Webber JL, Neware R, et al. Modified Lamport Merkle Digital Signature blockchain framework for authentication of internet of things healthcare data. Expert Systems 2022; 39(10): e12978. doi: 10.1111/exsy.12978
3. Abdullah S, Arshad J, Khan MM, Alazab M, Salah Ket al. PRISED tangle: A privacy-aware framework for smart healthcare data sharing using IOTA tangle. Complex & Intelligent Systems 2023; 9(3): 3023–3041. doi: 10.1007/s40747-021-00610-8
4. Shahid J, Ahmad R, Kiani AK, et al. Data protection and privacy of the internet of healthcare things (IoHTs). Applied Sciences 2022; 12(4): 1–22. doi: 10.3390/app12041927
5. Venkatesh S, Narasimhan K, Adalarasu K. An Overview of Interpretability Techniques for Explainable Artificial Intelligence (XAI) In Deep Learning-Based Medical Image Analysis, 2023 9th International Conference on Advanced Computing and Communication Systems (ICACCS); 17–18 March 2023; Coimbatore, India, 175–182. IEEE; 2023.
6. Borys K, Schmitt YA, Nauta M, et al. Explainable AI in medical imaging: An overview for clinical practitioners–Beyond saliency-based XAI approaches. European journal of radiology 2023; 1–11. doi: 10.1016/j.ejrad.2023.110786
7. Chai Y, Liu H, Xu J, et al. A multi-label classification with an adversarial-based denoising autoencoder for medical image annotation. ACM Transactions on Management Information Systems 2023; 14(2): 1–21. doi: 10.1145/3561653
8. Yashaswini S, Jayanthi MG, Subhash J. Captioning and Classification of Brain Tumor from MRI Images using Deep Learning Methods. Journal of Coastal Life Medicine 2023; 11: 302–307.
9. Jain K, Gandhi S, Singhal S, Rajput S. Semantic Image Captioning using Cosine Similarity Ranking with Semantic Search, Proceedings of the 2023 Fifteenth International Conference on Contemporary Computing 3–5 August 2023; Noida, India, 220–223. Association for Computing Machinery, New York, United States; 2023.
10. Gao D, Kong M, Zhao Y, et al. Simulating doctors' thinking logic for chest X-ray report generation via Transformer-based Semantic Query learning. Medical Image Analysis 2023; 102982. doi: 10.1016/j.media.2023.102982
11. Wong KK, Ayoub M, Cao Z, et al. The Synergy of Cybernetical Intelligence with Medical Image Analysis for Deep Medicine: A Methodological Perspective. Computer Methods and Programs in Biomedicine 2023; 107677. doi: 10.1016/j.cmpb.2023.107677
12. Yin Y, Han Z, Jian M, et al. AMSUnet: A neural network using atrous multi-scale convolution for medical image segmentation. Computers in Biology and Medicine 2023; 107120. doi: 10.1016/j.compbiomed.2023.107120
13. Dewi C, Chen RC, Yu H, Jiang X. XAI for Image Captioning using SHAP. Journal of Information Science & Engineering 2023; 39(4): 711–724. doi: 10.6688/JISE.202307_39(4).0001
14. Beddiar R, Oussalah M. Explainability in medical image captioning, Explainable Deep Learning AI, Academic Press; 2023. pp. 239–261.
15. Elbedwehy S, Medhat T, Hamza T, Alrahmawy MF. Enhanced descriptive captioning model for histopathological patches. Multimedia Tools and Applications 2023; 1–20. doi: 10.1007/s11042-023-15884-y

16. Lin Y, Lai K, Chang W. Skin Medical Image Captioning Using Multi-Label Classification and Siamese Network. IEEE Access 2023; 11: 23447–23454. doi: 10.1109/ACCESS.2023.3249462
17. Pang T, Li P, Zhao L. A survey on automatic generation of medical imaging reports based on deep learning. BioMedical Engineering OnLine 2023; 22(1): 1–16. doi: 10.1186/s12938-023-01113-y