Comparison of Fine-tuned and Frozen BiomedCLIP Encoders for Image Captioning of Cervical VIA Images with LSTM and BiLSTM Decoders

Tengku Kevin Juldianto[1*], Siti Nurmaini[2], Firdaus[3]

[1]Department of Computer Science, Faculty of Computer Science, Sriwijaya University, Palembang, South Sumatra, Indonesia.
[2,3]Intelligent System Research Group, Sriwijaya University, Palembang, South Sumatra, Indonesia.
[2,3]Artificial Intelligence Medical Center of Excellence, Sriwijaya University, Palembang, South Sumatra, Indonesia.
E-mail: [1*]tengkukevin05@gmail.com, [2]sitinurmaini@gmail.com, [3]firdaus@unsri.ac.id

(Received: 17 Sep 2026, revised: 27 Sep 2026, accepted: 30 Sep 2026)

Abstract

Visual Inspection with Acetic Acid (VIA) is a low-cost cervical cancer screening method, but its interpretation depends on the examiner and is often subjective. Existing artificial intelligence models for VIA images mostly output class labels or lesion areas without stating the visual findings behind the decision. This study applies image captioning to describe the findings in a VIA image as a sentence and examines which component most affects caption quality. Four configurations were built by combining two encoder treatments, a fine-tuned and a frozen BiomedCLIP encoder, with two decoders, Long Short-Term Memory (LSTM) and Bidirectional LSTM (BiLSTM). The dataset consists of 185 image–caption pairs from the IARC Image Bank VIA, split by stratification into 166 training and 19 test pairs. Caption quality was measured with BLEU, METEOR, ROUGE-L, and CIDEr, supported by the cosine similarity between test image features and the conclusion accuracy against a majority-class baseline. The fine-tuned encoder with the BiLSTM decoder achieved the highest scores, with a BLEU-4 of 74.98, METEOR of 83.77, ROUGE-L of 80.70, and CIDEr of 7.0212. The frozen encoder produced nearly uniform features, with a cosine similarity of 0.7634 compared with 0.2192 after fine-tuning, and generated only one or two distinct sentences across 19 test images. Its conclusion accuracy was 47.06%, below the baseline of 52.94%, whereas the fine-tuned encoder reached 70.59%. These results indicate that encoder fine-tuning is the main factor in captioning VIA images, while the choice between LSTM and BiLSTM has a smaller effect.

Keywords: BiLSTM, BiomedCLIP, Cervical Cancer, Image Captioning, Visual Inspection with Acetic Acid

I. INTRODUCTION

Cervical cancer is one of the leading causes of death among women worldwide [1], [2]. The onset of the disease is marked by lesions on the cervical surface, but these lesions are difficult to detect because they are located inside the body. Visual Inspection with Acetic Acid (VIA) is a method commonly used for the early detection of cervical cancer. VIA is performed by applying acetic acid to the cervical surface, after which a trained examiner observes the color changes in the cervical area [3]. Areas of the cervix that turn white (acetowhite) are considered suspected precancerous lesions. The presence of acetowhite does not always indicate a precancerous lesion, because squamous metaplasia, inflammation, and infection can also produce a similar white color [4]. A positive VIA result therefore still requires further examination. In addition, VIA interpretation relies entirely on the examiner's visual observation, so the result tends to be subjective and may differ between examiners for the same image [5]. Its accuracy and consistency also remain limited [6].

To reduce reliance on subjective judgment, a number of studies have applied artificial intelligence to VIA and cervicography images. In a systematic review, Viñals et al. [7] found that most of these efforts focus on classification tasks, namely categorizing images into normal and abnormal classes, or on segmentation tasks to mark lesion areas. The same review noted that while reported performance scores are generally high, the number of images used is limited and the data sources are not uniform across studies. Egemen et al. [8] highlighted this challenge from a clinical application

perspective, emphasizing that image-processing models for cervical screening must be evaluated for their suitability as a clinical test, from data collection methods and reference-labeling protocols to testing across different populations. Ahmed et al. [6] showed that this issue is real, demonstrating that model performance can change when the model is tested on a dataset different from its training data. In a context similar to the present study, Nurmaini et al. combined deep learning with medical guidelines to evaluate precancerous lesions from cervicography images captured before and after acetic acid application [5], and developed a system to support the diagnosis of cervical dysplasia from cervicography images [9]. Based on these previous studies, the output of deep learning models generally consists of either a class label or an area marked as a precancerous lesion. A label does provide a decision, but it does not state which visual features underlie that decision, making it difficult for healthcare professionals to trace back the basis of the model's assessment [10].

Image captioning models offer a different form of output by receiving an image as input and generating a sentence that describes its content [11]. These models have been applied to several types of medical images. For instance, they are used to draft examination reports in radiology [12] and to generate medical reports directly from skin lesion images [13]. Similar approaches have also been developed to draft reports and at the same time answer questions about the content of medical images [14]. Unlike single-label classification models, the generated descriptions contain more information, such as lesion location, margin characteristics, and the size of the affected area [15], [16]. This resembles how medical experts document their findings, making the results easier for humans to review. In Indonesia, a CNN–LSTM architecture was used to construct a dataset alongside a captioning model [17], showing that a captioning workflow can be arranged for a new dataset of limited size. However, to the best of the authors' knowledge, image captioning has not yet been applied to cervical VIA images. This absence is notable because VIA findings are inherently descriptive: examiners report several visual characteristics, such as squamocolumnar junction visibility, the presence, thickness, margin, and surface of acetowhite lesions, and their size and location, before reaching a conclusion [3]. A model that generates such descriptions would match the way VIA results are documented, rather than reducing them to a single label.

An image captioning system generally consists of an encoder and a decoder [18]. The encoder converts an image into a representation vector, often referred to as a feature or embedding. The decoder then translates that vector into a descriptive sentence [19]. The encoder used in this study is BiomedCLIP, a domain-specific variant of the CLIP (Contrastive Language–Image Pre-training) architecture. CLIP is trained on image–text pairs by bringing matching image and text features closer in the representation space while separating non-matching pairs, an approach known as contrastive learning [20]. Because these pre-trained representations can be transferred to downstream tasks without training from scratch, CLIP-based models have been extended to the medical domain [21]. BiomedCLIP follows the same training approach as CLIP, using fifteen million biomedical image–text pairs [22]. With such training data, the embeddings generated by BiomedCLIP are expected to be better suited for medical images than those from standard CLIP, which is trained on general-domain images. The numerical features generated by the encoder are then fed into the decoder for caption generation.

The first decoder used is Long Short-Term Memory (LSTM), a type of Recurrent Neural Network that generates sentences word by word. LSTM retains information in a memory cell, maintaining long-range dependencies between words [23]. In the form commonly used today, LSTM incorporates three gates, namely the input, forget, and output gates, which regulate which information is retained or discarded [24]. Combining an image encoder with an LSTM decoder remains widely used in medical image captioning [25]. The second decoder is Bidirectional LSTM (BiLSTM), which processes a sequence in both the forward and backward directions and combines the two representations [26]. This bidirectional mechanism has been reported to be beneficial for captioning tasks [27] and for other tasks involving sequential data [28]. During caption generation, however, the words after the current position have not yet been produced, so the BiLSTM reads only the partial sentence generated so far, once from left to right and once from right to left. This study does not combine the two decoders in a single model; it compares them as alternative decoders under the same conditions. Each option involves a trade-off. LSTM has fewer parameters and follows the natural left-to-right order of word generation, which suits a small dataset, but it summarizes the partial sentence from only one direction. BiLSTM captures the relationships between words in the partial sentence more fully, but it doubles the number of recurrent parameters, which increases the risk of overfitting when training data are limited. Because sentences describing VIA findings have a relatively fixed structure, it is not obvious whether the additional reading direction improves word selection enough to offset this cost, which motivates the comparison.

Besides the decoder architecture, captioning performance is also influenced by the encoder strategy. The encoder weights can be fine-tuned using new data (fine-tuning) or kept fixed without retraining (freezing). A review of transfer learning on medical images shows that this choice has a substantial impact, especially when training data are limited [29]. A direct comparison between a fine-tuned and a frozen encoder has been conducted on other medical imaging modalities, such as Gram-stained bacterial microscopy images [30]; however, that study evaluated a classification task rather than sentence generation. Two gaps therefore remain. First, image captioning has not been applied to cervical VIA images. Second, existing comparisons of fine-tuned and frozen encoders on medical images address classification [29], [30], so it is unclear whether the same conclusion holds for sentence generation, where the decoder must turn image features into a sequence of clinical findings. The encoder treatment and the decoder type need to be examined together because the decoder can only produce varied and correct sentences if the encoder distinguishes the images, so the effect of one component

DOI: 10.34148/teknika.v15i3.1593

cannot be assessed without the other. To the best of the authors' knowledge, no study has evaluated this combination for image captioning on cervical VIA images, and this gap forms the basis of the present study.

This study examines four system combinations based on two factors: encoder treatment (fine-tuning or freezing) and decoder type (LSTM or BiLSTM). The study has three objectives. The first is to compare the performance of these four combinations on cervical VIA images. The second is to examine the effect of encoder treatment on the generated sentences. The third is to assess the quality of the generated sentences from two perspectives: their similarity to the reference sentences and the correctness of the conclusion they contain.

The evaluation metrics for the generated sentences are BLEU [31], METEOR [32], ROUGE-L [33], and CIDEr [34]. The diversity of the sentences across the test images is also examined, because captioning models tend to produce generic sentences that are similar to one another [35]. The correctness of the conclusion is examined through the accuracy of the positive or negative conclusion, which is then compared with a majority class baseline. An assessment from these two perspectives is needed because automatic scores on medical text generation tasks have been reported not to always align with the correctness of the report content [36]. The number of images available in this study is limited, so the results should be interpreted as a comparison between combinations under limited data conditions, rather than as a claim of generalizable performance [37].

II. RESEARCH METHOD

This study is conducted through seven sequential stages, as shown in Figure 1.

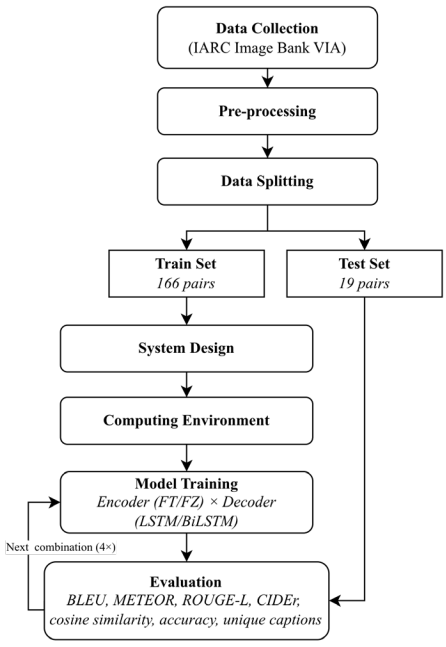


Figure 1. Research Method

The model training and evaluation stages are repeated for each of the four configurations, as detailed in Algorithm 1. The first stage is data collection, which involves preparing VIA images along with their examination result notes from the data source used. The second stage is pre-processing, which involves preparing the images to match the encoder's input size and organizing the examination result notes into caption sentences. The third stage is data splitting into training and testing data. The fourth stage is system design, which involves determining the structure of the encoder and decoder along with the size of each layer. The fifth stage is the setup of the computing environment and its supporting tools used for training and evaluation. The sixth stage is model training, which is performed in two parts: training the encoder in the fine-tune configuration and training the decoder to generate captions. The seventh stage is evaluation, which involves measuring the quality of the generated captions using predefined metrics.

A. Data Collection

The data are taken from the IARC Image Bank VIA [3]. This dataset contains 186 VIA screening cases. Each case consists of cervical images before acetic acid application, images after acetic acid application, and screening outcome notes. After sorting the files, one case has no image (case 166), leaving 185 cases to be used. Each case provides several color photographs of the same cervix taken at different examination steps. The images are stored in JPG format with an original resolution of 920 × 720 pixels. Only the image taken after acetic acid application is used in this study, as explained in the pre-processing stage. The screening outcome notes are stored in two files, as shown in Table 1.

Table 1. Completeness of IARC Data

File Name	Description
Cases – Images.xlsx	Descriptions of each image: before acetic acid application, after acetic acid application, after acetic acid application with a magnified view, and after iodine application.
IARCImageBankVIA Cases Meta data.xlsx	Columns of findings along with their descriptions according to VIA standards.

The metadata file contains clinical parameters following VIA reporting standards, including squamocolumnar junction (SCJ) visibility, the presence of acetowhite lesions, the thickness of the acetowhite area, margin shape, surface condition, lesion size and location, as well as the examination conclusion. These parameters are arranged into captions during the pre-processing stage. VIA examination conclusions consist of three categories: Negative means no acetowhite lesions are found on the cervical surface; Positive means acetowhite lesions indicating suspected precancerous lesions are found; Suspicious of cancer means features pointing to potential cancer are found, requiring further examination. Because each caption is composed from the clinical parameters

recorded for its case, caption length depends on the findings. A positive case, for example, is described as "the scj is fully visible located on ectocervix acetowhite lesion present with white area thin margin sharp but irregular surface smooth located within tz or close to the external os if scj is not visible size covering less than 75 of ectocervix via result positive", whereas a negative case is described briefly as "no acetowhite lesion observed via result negative". All captions share the same limited set of clinical terms, which explains the small vocabulary reported in the Data Split section. The number of cases in each category is shown in Table 2.

Table 2. Number of Cases per Screening Outcome

Screening Outcome	Number of Cases
Negative	92
Positive	73
Suspicious of cancer	20
Total	185

B. Pre-Processing

Pre-processing is performed on two sides: the image side and the text side. On the image side, only the image labeled "after application of acetic acid" is taken from each case, as this image corresponds to the VIA procedure. The images are then cropped and resized to 224 × 224 pixels to match the input size of BiomedCLIP. On the text side, clinical parameters in the metadata file are first composed into a single medical sentence for each case. This sentence is then cleaned through four steps: converting all text to lowercase, removing apostrophes, replacing non-alphanumeric characters with spaces, and trimming excessive spaces. Character filtering intentionally retains numbers, as lesion size descriptions like "covering less than 75%" would lose their meaning if the numbers were discarded. Finally, each sentence is given start and end marker tokens, startseq and endseq, to signal the decoder when to start and end sentence generation. After completing this pre-processing stage, a dataset of 185 image–caption pairs is produced.

C. Data Split

The data are split into approximately 90% training data (166 of 185 pairs, 89.73%) and 10% test data (19 of 185 pairs, 10.27%). The total amount of data in this study is only 185 pairs, so all data outside the test set is used to train the model. The split uses stratification based on the final examination outcome, namely negative, positive, and suspicious of cancer, so that the proportions of these three categories remain balanced between the training and test data. The seed value is set to a fixed number so that the split is reproducible. Out of the 185 pairs, 166 pairs become training data and 19 pairs become test data. The details are presented in Table 3.

Data splitting is performed prior to fine-tuning the encoder so that the encoder never sees the test data [37]. For the same reason, the tokenizer and the maximum caption length are constructed exclusively from the training data. From the training data, a maximum caption length of 52 tokens and a vocabulary size of 57 words are obtained. This small vocabulary size is a characteristic of this dataset, as all captions are constructed from the same clinical parameters with a limited choice of words.

Table 3. Distribution of Training and Test Data

Category	Train Set	Test Set	Total
Negative	83	9	92
Positive	65	8	73
Suspicious of cancer	18	2	20
Total	166	19	185
Percentage	89.73%	10.27%	100%

D. System Design

The constructed system takes a single VIA image as input and generates a single descriptive sentence. The sentence is generated one word at a time. The image captioning system is divided into two components: the encoder, which converts the image into a sequence of numbers representing its content, and the decoder, which transforms that sequence of numbers into words. The complete model architecture is shown in Figure 2, and each block is numbered according to the explanation below.

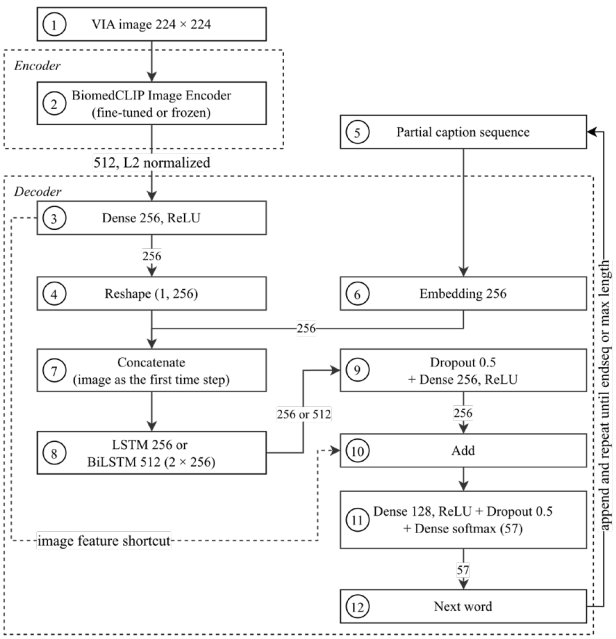


Figure 2. Model Architecture

In the encoder component, the image (1) is processed by BiomedCLIP (2) and converted into a 512-dimensional embedding, which is normalized using the L2 norm so that all features reside on the same scale. In the decoder, this embedding is first projected into 256 dimensions through a Dense layer with a ReLU activation function (3). This projection is necessary because each word in the decoder is also represented by 256 numerical values. With matching dimensions, both image features and word features can be treated as a single input type.

The 256-dimensional image features are reshaped into a single time step (4). At the same time, the partial caption

generated so far (5) is converted into 256-dimensional word vectors by an embedding layer (6). The image features are then placed in front of the sequence of words through concatenation (7). With this arrangement, the image is treated as the first word of the sentence to be constructed. The decoder reads the image first, followed by the words that have already been generated. This method ensures the decoder always knows which image is being described at each step of sentence generation.

The sequence is processed by the decoder (8), which is either an LSTM with 256 units or a BiLSTM that produces a 512-dimensional output because it reads the sequence from both directions. The decoder output passes through a dropout layer with a rate of 0.5 to reduce overfitting, then a Dense layer with 256 units and ReLU activation (9).

This result is added back to the projected image features through the image feature shortcut (10). This addition is necessary because the image features only enter once at the beginning of the sequence, causing their influence to fade after passing through multiple steps in the decoder. With this re-addition, the image information is preserved up to the final layer. After that, there is a Dense layer with 128 units and ReLU activation, a second dropout of 0.5, and a Dense output layer with a softmax activation function matching the vocabulary size (11). The softmax output contains probabilities for each word, and the word with the highest probability is selected as the next word (12). This word is appended to the partial caption (5), and the process is repeated until the model generates an end-of-sentence marker token.

Algorithm 1. Research procedure for the four encoder–decoder combinations

Input: 185 VIA image–caption pairs
Output: evaluation scores of the four combinations (Table 7)

1: Pre-process images (224 × 224) and captions (add startseq, endseq)
2: Split into 166 training and 19 test pairs (stratified, seed = 42)
3: Build tokenizer and maximum caption length L from training data
4: **for each** encoder $E \in$ {fine-tuned, frozen} **do**
5: **if** E = fine-tuned **then** fine-tune visual tower (contrastive loss)
6: Extract 512-dimensional image embeddings with E
7: **for each** decoder $M \in$ {LSTM, BiLSTM} **do**
8: Train M on training embeddings (teacher forcing)
9: **for each** test image x **do**
10: *sequence* ← [startseq]
11: **repeat** append the most probable next word w to *sequence*
12: **until** w = endseq **or** length(*sequence*) = L
13: **end for**
14: Compute BLEU, METEOR, ROUGE-L, CIDEr, cosine similarity, accuracy, and unique captions
15: **end for**
16: **end for**

The encoder is tested under two treatments, fine-tune and freeze, each paired with LSTM and BiLSTM, resulting in four configurations. The four configurations are not executed simultaneously. They are run one at a time with the same procedure: the embeddings are extracted once for each encoder treatment, both decoders are then trained on the same embeddings, and the results of each configuration are collected into Table 7. Within each configuration, caption generation is also iterative: each predicted word is appended to the partial caption and fed back into the decoder until the end marker is produced, as shown by the loop in Figure 2 and in Algorithm 1.

All four use the same data split, seed, and hyperparameters so that any difference in results comes solely from the encoder treatment and decoder type.

E. Computing Environment

The entire training and evaluation process is conducted on Google Colab with a Pro plan. The resource specifications used are shown in Table 4. Additionally, Modern CSV is used to view dataset CSV files, and Python scripts are used to combine clinical parameters into captions.

Table 4. Computing Environment Specifications

Component	Description
GPU	A100
RAM	84 GB
GPU RAM	40 GB
Disk	235 GB

F. Model Training

This study focuses on the combination of encoder treatment and decoder type, resulting in four training combinations as shown in Table 5.

Table 5. Model Combinations

Name	Encoder	Decoder
Combination 1	BiomedCLIP (FT)	LSTM
Combination 2	BiomedCLIP (FT)	BiLSTM
Combination 3	BiomedCLIP (FZ)	LSTM
Combination 4	BiomedCLIP (FZ)	BiLSTM

In the fine-tune configuration, the BiomedCLIP encoder is also trained using VIA images so that the model recognizes the features within those images. The freeze configuration is used as a comparison to observe the effect of fine-tuning. Training is divided into two stages: encoder training and decoder training. Before training begins, the entire process is fixed with a seed value of 42 to ensure reproducible results when the program is re-run.

The first stage is encoder training, which is only executed in the fine-tune configuration (Algorithm 1, line 5). The encoder component that reads images (visual tower) is fine-tuned using VIA images via contrastive learning; specifically, the model is given image–text pairs in each batch and learns to bring matching image and text representations closer while pushing non-matching ones apart [20]. The difference is calculated using contrastive loss with a temperature value of 0.07. At this stage, only the image-reading component is

trained, while the text component remains fixed. Training uses only the training data. The optimizer is AdamW with a learning rate of 5×10^{-5}, a weight decay of 1×10^{-4}, a batch size of 16, and 20 epochs. In the freeze configuration, this stage is skipped, so BiomedCLIP is used as-is.

Once the encoder is ready (line 6), all images are converted into 512-dimensional embeddings in inference mode without gradient computation. This extraction includes test images, but it does not constitute data leakage because the encoder weights have only ever been trained using the training data. These embeddings serve as the input for the decoder.

The second stage is training the decoder (line 8), both LSTM and BiLSTM, to generate captions word by word. The training utilizes teacher forcing, where the decoder is provided with the correct word from the reference caption to predict the next word. Each caption is therefore split into several training pairs, consisting of a partial sentence as input and the subsequent word as the target. The length of each partial sentence is then standardized through padding based on the longest caption in the training data. The loss function used is categorical cross-entropy with the Adam optimizer. Training is executed for a maximum of 100 epochs with a batch size of 64.

Since there is no validation data, the entire training control mechanism monitors the training data loss. ModelCheckpoint saves the weights with the best loss. EarlyStopping halts training if the loss does not improve for 5 epochs and restores the best weights. ReduceLROnPlateau halves the learning rate if the loss remains unchanged for 3 epochs, with a lower bound of 1×10^{-6}. A dropout of 0.5 is applied in two locations to mitigate overfitting. All hyperparameters are summarized in Table 6.

Table 6. Training Hyperparameters

Stage	Parameter	Value
Encoder (FT)	Optimizer	AdamW
	Learning Rate	5×10^{-5}
	Weight decay	1×10^{-4}
	Batch size	16
	Epoch	20
	Temperature	0.07
Decoder	Optimizer	Adam
	Loss	Categorical cross-entropy
	Batch size	64
	Epoch (Max)	100
	Dropout	0.5
	Monitor callback	Train loss
	Early Stop (patience)	5
	ReduceLROnPlateau (factor / patience)	0.5 / 3
	Minimum LR	1×10^{-6}
	Embedding dimension	256
	LSTM / BiLSTM units	256

G. Evaluation

Sentences are constructed using greedy decoding, where at each step, a single word with the highest probability from the softmax output is taken, without beam search. Generation stops when the endseq token appears or when the sentence length reaches the maximum limit.

The quality of the sentences generated by the model is measured using four metrics commonly used in image captioning: BLEU, METEOR, ROUGE-L, and CIDEr. All four values are obtained by comparing the model's captions against the reference (ground truth) captions across 19 test images. BLEU is calculated using the NLTK library with the corpus_bleu function along with smoothing, as the small number of test images causes some n-grams to never match, resulting in a zero score without smoothing. METEOR is also calculated using NLTK, while ROUGE-L and CIDEr are calculated using the pycocoevalcap library. The higher the value, the more similar the model's caption is to the reference caption.

Additionally, supporting measures are calculated. Cosine similarity between test image features is used to check whether the features of the test images are nearly uniform. The accuracy of positive/negative conclusions is used to observe the alignment between the conclusions in the model's sentences and the actual conclusions. The number of unique captions is also recorded to observe the diversity of the generated captions [35].

BLEU (Bilingual Evaluation Understudy) measures the similarity of word fragments (n-grams) between the model sentence and the reference sentence, as formulated in Equation (1) and (2) [31].

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \quad (1)$$

$$BP = \begin{cases} 1, & if\ c > r \\ \exp\left(1 - \frac{r}{c}\right), & if\ c \le r \end{cases} \quad (2)$$

Here p_n is the n-gram precision, which represents how many of the model's n-grams appear in the reference sentence. w_n is the weight of each n-gram, which equals $1/N$. BP is the brevity penalty applied when the model sentence is too short, where c is the model sentence length and r is the reference sentence length. METEOR (Metric for Evaluation of Translation with Explicit ORdering) evaluates word alignment while accounting for missing words and word order, as formulated in Equation (3), (4), and (5) [32].

$$METEOR = F_{mean} \cdot (1 - Pen) \quad (3)$$

$$F_{mean} = \frac{10\,P\,R}{R + 9P} \quad (4)$$

$$Pen = \gamma \left(\frac{ch}{m}\right)^{\beta} \quad (5)$$

DOI: 10.34148/teknika.v15i3.1593

CC BY SA

Here F_{mean} is the weighted harmonic mean of unigram precision and recall, while Pen is the fragmentation penalty that accounts for word order. P is precision and R is recall at the word level. ch is the number of matching consecutive word groups (chunks), m is the number of matching words, while γ and β are penalty parameters, set to 0.5 and 3 respectively following the default values in the NLTK implementation. ROUGE-L (Recall-Oriented Understudy for Gisting Evaluation – Longest Common Subsequence) evaluates similarity based on the longest common word sequence between the model sentence and the reference sentence, and is formulated in Equation (6), (7), and (8) [33].

$$R_{lcs} = \frac{LCS(X,Y)}{m} \tag{6}$$

$$P_{lcs} = \frac{LCS(X,Y)}{n} \tag{7}$$

$$F_{lcs} = \frac{(1+\beta^2)\, R_{lcs}\, P_{lcs}}{R_{lcs} + \beta^2 P_{lcs}} \tag{8}$$

Here $LCS(X,Y)$ is the length of the longest common word sequence between the reference sentence X and the model sentence Y. m is the length of the reference sentence, n is the length of the model sentence, and β is a weighting parameter, which is unrelated to the β used in Equation (5). CIDEr (Consensus-based Image Description Evaluation) gives more weight to rarely occurring terms, such as medical terms, by calculating the similarity between TF-IDF weighted n-grams, and is formulated in Equation (9) and (10) [34].

$$CIDEr_n(c,S) = \frac{1}{M}\sum_{i=1}^{M} \frac{g^n(c)\cdot g^n(s_i)}{\| g^n(c) \| \| g^n(s_i) \|} \tag{9}$$

$$CIDEr = \sum_{n=1}^{N} w_n\, CIDEr_n \tag{10}$$

Here $g^{\mathrm{n}}(\cdot)$ is the TF-IDF weight vector of the n-gram, c is the model sentence, s_{i} is the i-th reference sentence, and M is the number of reference sentences. N is the maximum n-gram length, which is 4, and w_n is the weight for each n-gram length, which equals $1/N$. The first supporting measure is cosine similarity, which is a measure of similarity between test image features. This quantity is the same measure that CLIP optimizes when training the encoder, namely the dot product between normalized features, as formulated in Equation (11) [20].

$$cosine_similarity(u,v) = \frac{u\cdot v}{\| u \| \| v \|} \tag{11}$$

Here u and v are the two image feature vectors being compared. The calculation is performed for all test image pairs, and then all values are averaged into a single number. The second supporting measure is the accuracy of the positive/negative conclusion, which is the alignment between the conclusion in the model sentence and the true conclusion. Accuracy is formulated in Equation (12).

$$Accuracy = \frac{\sum correct\ predictions}{\sum clearly\ labeled\ test\ data} \times 100\% \tag{12}$$

Clearly labeled test data refers to images with a positive or negative conclusion, while the category suspicious of cancer is excluded because it cannot be mapped to either of the two. As an accuracy benchmark, a majority class baseline is used, which is the accuracy obtained if the model simply predicts the category that appears most frequently.

III. RESULT AND DISCUSSION

The results of the four configurations are shown in Table 7. The test data consist of 19 images, comprising 9 images with a negative conclusion, 8 with a positive conclusion, and 2 with a suspicious of cancer conclusion. The accuracy calculation only uses 17 clearly labeled images, which are those with a positive or negative conclusion. The majority class in this group is negative with 9 images, resulting in a majority class baseline value of 52.94%.

Table 7. Evaluation Scores of Encoder and Decoder Combinations

Metric	FT-LSTM	FT-BiLSTM	FZ-LSTM	FZ-BiLSTM
BLEU-1	74.73	79.23	66.90	66.90
BLEU-2	73.10	77.74	64.82	64.88
BLEU-3	71.45	76.24	62.90	62.99
BLEU-4	70.05	74.98	61.51	61.57
METEOR	80.17	83.77	75.75	75.76
ROUGE-L	76.51	80.70	70.29	70.41
CIDEr	6.1640	7.0212	5.0655	5.0376
Cosine similarity	0.2192	0.2192	0.7634	0.7634
Accuracy	70.59%	70.59%	47.06%	47.06%
Baseline	52.94%	52.94%	52.94%	52.94%
Unique captions	7/19	8/19	1/19	2/19

Each combination shows a distinct pattern. Combination 1 (FT-LSTM) reaches a BLEU-4 of 70.05, METEOR of 80.17, ROUGE-L of 76.51, and CIDEr of 6.1640, and produces 7 unique captions for 19 test images. Combination 2 (FT-BiLSTM) obtains the highest scores on all metrics, with a BLEU-4 of 74.98, METEOR of 83.77, ROUGE-L of 80.70, and CIDEr of 7.0212, and produces 8 unique captions. Both fine-tuned combinations reach a conclusion accuracy of 70.59%, above the baseline of 52.94%. Combination 3 (FZ-LSTM) reaches a BLEU-4 of 61.51, METEOR of 75.75, ROUGE-L of 70.29, and CIDEr of 5.0655, but produces only one caption for all test images. Combination 4 (FZ-BiLSTM) gives almost the same scores, with a BLEU-4 of 61.57, METEOR of 75.76, ROUGE-L of 70.41, and CIDEr of 5.0376,

and produces two unique captions. Both frozen combinations reach a conclusion accuracy of 47.06%, below the baseline. The following paragraphs discuss the factors behind these differences. The treatment applied to the encoder has a much greater impact than the type of decoder used. All configurations with a fine-tuned encoder achieve higher scores across every metric compared to configurations with a frozen encoder. The difference in BLEU-4 reaches 13.41 points between FT-BiLSTM and FZ-BiLSTM, while the ROUGE-L difference reaches 10.29 points. Differences are also visible in CIDEr, which is 7.0212 for FT-BiLSTM compared to 5.0376 for FZ-BiLSTM. CIDEr assigns greater weight to rare n-grams and penalizes overly common n-grams [34], making this metric the most sensitive to repeated sentences. Similar results have been reported on other types of medical images, where fine-tuned encoders yield better results than frozen encoders [29], [30].

The cause of this difference can be observed in the cosine similarity between the features of the test images. In the configuration with a frozen encoder, the value reaches 0.7634, whereas with the fine-tuned encoder, it is only 0.2192. A value approaching one indicates that the features across all test images point in nearly the same direction, meaning the encoder practically does not distinguish one image from another. This value remains identical for both decoder types since cosine similarity depends solely on the encoder. The direct consequence is visible in the output: the frozen encoder configuration generates only one to two sentences across all 19 test images, while the fine-tuned encoder configuration produces 7 to 8 distinct sentences. BiomedCLIP used without fine-tuning does not recognize the distinctive characteristics of VIA images; as a result, the decoder receives nearly uniform input and lacks the foundation to generate varied sentences. The tendency of models to produce uniform sentences regardless of differences between images has also been reported in research on sentence diversity in image captioning [35].

The clinical impact of this uniformity is evident in the accuracy of the conclusions. The configuration with the fine-tuned encoder achieves an accuracy of 70.59%, correctly concluding 12 out of 17 clearly labeled images, thereby surpassing the majority-class baseline of 52.94%. Conversely, the configuration with the frozen encoder only achieves 47.06%, or 8 correct conclusions, falling below the baseline. These figures can be explained by the model's behavior. Because it produces only a single sentence for all images and that sentence concludes positive, the model predicts positive for every image. Its accuracy is therefore simply equal to the proportion of positive images in the test set, which is 8 out of 17. A model that consistently predicts negative would actually be correct for 9 out of 17 images. In other words, the configuration with the frozen encoder performs worse than majority-class guessing, even though its BLEU score reaches 61.57.

On the fine-tuned encoder, BiLSTM achieves higher scores than LSTM across all metrics. The achieved BLEU-4 is 74.98 compared to 70.05, METEOR is 83.77 compared to 80.17, ROUGE-L is 80.70 compared to 76.51, and CIDEr is 7.0212 compared to 6.1640. BiLSTM also generates slightly more unique captions, namely 8 compared to 7. However, the conclusion accuracy for both decoders is identical at 70.59%, indicating that BiLSTM's advantage lies in sentence similarity to the reference rather than the correctness of the conclusions. On the frozen encoder, the difference between the two decoders nearly disappears, with a BLEU-4 difference of only 0.06 points and identical accuracy. This indicates that the decoder's advantage only manifests when the received image features are sufficiently distinguishable.

The results in Table 7 also highlight the limitations of automated metrics. The FZ-LSTM configuration achieves a METEOR score of 75.75, which is only 4.42 points below FT-LSTM's score of 80.17, even though FZ-LSTM generates only a single sentence for all test images and its conclusion accuracy falls below the baseline. This occurs because all captions in this dataset are constructed from the same clinical parameters with a limited vocabulary, allowing even a single generic sentence to share high word overlap with the reference sentence. Consequently, metrics measuring word matching exhibit a high lower bound on datasets of this nature. A similar finding has been reported in automated radiology reporting, where automated scores do not always align with the correctness of the report content [36]. This finding underscores that evaluating medical image captioning systems cannot rely solely on sentence similarity scores, but must also incorporate an assessment of the correctness of the conclusions.

Sample test results for the best configuration are presented in Table 8. The model is capable of generating structured medical sentences that follow the reference caption patterns, such as identifying the visibility of the squamocolumnar junction, the presence or absence of acetowhite lesions, and the characteristics of the lesion's margins and surface.

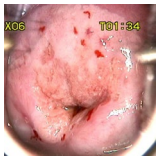
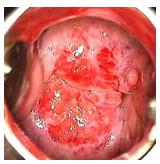
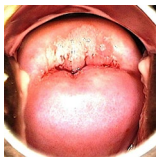
In the cases of ADS1, ADX1, and AFA1, the generated sentences are identical to the reference sentences, including the conclusion. In the AAS1 case, the generated conclusion is correct, but the lesion margin feature is incorrect, being stated as sharp and regular when it should be diffuse. The largest error occurs in the ABK1 case, which is an image with a suspicious of cancer conclusion that is declared positive by the model, accompanied by errors in the visibility of the squamocolumnar junction and the lesion size. This case demonstrates the model's limitations in the class with the smallest amount of data.

The proposed approach offers three advantages. First, the output states each finding separately, such as SCJ visibility, acetowhite characteristics, lesion size and location, and the conclusion, so an examiner can verify each finding against the image rather than receiving only a class label. Second, because the captions follow the structure of VIA reporting standards, the output can serve as a consistent draft of the examination record. Third, only the visual component of a domain-specific encoder is fine-tuned and the decoder is lightweight, so the system can be trained on a small dataset with a single GPU. Its disadvantages are that the captions are limited to a fixed template with a 57-word vocabulary, the images come from a single source, so performance on images from other devices and lighting conditions is unknown, and a conclusion accuracy

DOI: 10.34148/teknika.v15i3.1593

of 70.59% is not sufficient for independent use. In practice, VIA is often performed by midwives and nurses in primary health care facilities with limited specialist support [3], [4]. In such settings, this system could act as a second reader that drafts structured findings for the examiner to verify, or as a training aid for health workers learning to interpret VIA images. It is not intended to replace the examiner, and clinical validation on local screening data is required before deployment.

Table 8. Model Output

Case	Output	Ground Truth
 ADS1	the scj is fully visible located on ectocervix acetowhite lesion present with white area thin margin sharp but irregular surface smooth located within tz or close to the external os if scj is not visible size covering less than 75 of ectocervix via result positive	the scj is fully visible located on ectocervix acetowhite lesion present with white area thin margin sharp but irregular surface smooth located within tz or close to the external os if scj is not visible size covering less than 75 of ectocervix via result positive
 ABK1	the scj is fully visible located on ectocervix acetowhite lesion present with white area thin margin sharp and regular surface smooth located within tz or close to the external os if scj is not visible size covering less than 75 of ectocervix via result positive	the scj is not visible acetowhite lesion present with white area dense margin sharp and regular surface irregular located within tz or close to the external os if scj is not visible size covering more than 75 of ectocervix via result suspicious of cancer
 AAS1	the scj is fully visible located on ectocervix acetowhite lesion present with white area thin margin sharp and regular surface smooth located within tz or close to the external os if scj is not visible size covering less than 75 of ectocervix via result negative	the scj is fully visible located on ectocervix acetowhite lesion present with white area thin margin diffuse surface smooth located within tz or close to the external os if scj is not visible size covering less than 75 of ectocervix via result negative
 AFA1	no acetowhite lesion observed via result negative	no acetowhite lesion observed via result negative
 ADX1	the scj is fully visible located on ectocervix acetowhite lesion present with white area thin margin sharp but irregular surface smooth located within tz or close to the external os if scj is not visible size covering less than 75 of ectocervix via result positive	the scj is fully visible located on ectocervix acetowhite lesion present with white area thin margin sharp but irregular surface smooth located within tz or close to the external os if scj is not visible size covering less than 75 of ectocervix via result positive

This study has several limitations. First, the results are obtained from a single training run without repetition, so variance measures across trials are not yet available [37]. Second, the test set consists of only 19 images, with only 17 of them used in accuracy calculations, making the accuracy score unstable against changes in data composition. Third, the suspicious of cancer class contains only 20 images in the entire dataset and 2 images in the test set, and is excluded from the accuracy calculation, meaning the model's performance on this class remains unmeasured. Fourth, retraining the encoder is not fully deterministic, so scores under the fine-tune configuration may vary between runs.

IV. CONCLUSION

This study compares four combinations of image captioning model architectures on VIA images, using both fine-tuned and frozen encoders, each paired with LSTM and BiLSTM decoders. The results indicate that fine-tuning the encoder is the primary determinant of system success. Configurations with a fine-tuned encoder achieve the highest scores across all metrics, with the fine-tuned encoder and BiLSTM decoder pair reaching a BLEU-4 of 74.98, METEOR of 83.77, ROUGE-L of 80.70, and CIDEr of 7.0212. The frozen encoder produces nearly uniform features across all test images, as indicated by a cosine similarity of 0.7634 compared to 0.2192 for the fine-tuned encoder, causing the model to generate only one to two distinct sentences across 19 test images.

In terms of conclusion accuracy, the configuration with the fine-tuned encoder achieves an accuracy of 70.59%, surpassing the majority-class baseline of 52.94%. Conversely, the configuration with the frozen encoder only reaches 47.06%, falling below the baseline. On the decoder side,

BiLSTM achieves higher sentence similarity scores than LSTM when the encoder is fine-tuned, but the conclusion accuracy for both is identical, and the performance gap nearly disappears when the encoder is frozen.

The amount of data used is indeed limited, consisting of 185 image–caption pairs, with 166 pairs used as training data; thus, model quality could still be improved with a significantly larger medical dataset. For future research, it is recommended to test across multiple random seeds and report the mean and standard deviation, expand the training dataset, specifically address the suspicious of cancer class, and involve medical professionals to evaluate caption quality.

REFERENCES

[1] F. Bray et al., "Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA Cancer J. Clin.*, vol. 74, no. 3, pp. 229–263, 2024, doi: 10.3322/caac.21834.

[2] D. Singh et al., "Global estimates of incidence and mortality of cervical cancer in 2020: a baseline analysis of the WHO Global Cervical Cancer Elimination Initiative," *Lancet Glob. Health*, vol. 11, no. 2, pp. e197–e206, 2023, doi: 10.1016/S2214-109X(22)00501-0.

[3] S. Mittal and P. Basu, *Atlas of Visual Inspection of the Cervix with Acetic Acid for Screening, Triage, and Assessment for Treatment (IARC CancerBase No. 16)*. Lyon, France: International Agency for Research on Cancer, 2020. [Online]. Available: https://screening.iarc.fr/atlasvia.php

[4] IARC Working Group on the Evaluation of Cancer-Preventive Interventions, *Cervical Cancer Screening (IARC Handbooks of Cancer Prevention, vol. 18)*. Lyon, France: International Agency for Research on Cancer, 2022.

[5] S. Nurmaini et al., "Robust assessment of cervical precancerous lesions from pre- and post-acetic acid cervicography by combining deep learning and medical guidelines," *Inform. Med. Unlocked*, vol. 52, Art. no. 101609, 2025, doi: 10.1016/j.imu.2024.101609.

[6] S. R. Ahmed et al., "Assessing generalizability of an AI-based visual test for cervical cancer screening," *PLOS Digit. Health*, vol. 3, no. 10, Art. no. e0000364, 2024, doi: 10.1371/journal.pdig.0000364.

[7] R. Viñals, M. Jonnalagedda, P. Petignat, J.-P. Thiran, and P. Vassilakos, "Artificial intelligence-based cervical cancer screening on images taken during visual inspection with acetic acid: A systematic review," *Diagnostics*, vol. 13, no. 5, Art. no. 836, 2023, doi: 10.3390/diagnostics13050836.

[8] D. Egemen et al., "Artificial intelligence–based image analysis in clinical testing: lessons from cervical cancer screening," *J. Natl. Cancer Inst.*, vol. 116, no. 1, pp. 26–33, 2024, doi: 10.1093/jnci/djad202.

[9] S. Nurmaini et al., "AI-assisted diagnosis of cervical dysplasia from cervicography images," *Sci. Rep.*, vol. 16, Art. no. 9920, 2026, doi: 10.1038/s41598-026-39192-1.

[10] M. Ghassemi, L. Oakden-Rayner, and A. L. Beam, "The false hope of current approaches to explainable artificial intelligence in health care," *Lancet Digit. Health*, vol. 3, no. 11, pp. e745–e750, 2021, doi: 10.1016/S2589-7500(21)00208-9.

[11] M. Stefanini, M. Cornia, L. Baraldi, S. Cascianelli, G. Fiameni, and R. Cucchiara, "From show to tell: A survey on deep learning-based image captioning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 539–559, Jan. 2023, doi: 10.1109/TPAMI.2022.3148210.

[12] P. Messina et al., "A survey on deep learning and explainability for automatic report generation from medical images," *ACM Comput. Surv.*, vol. 54, no. 10s, pp. 1–40, 2022, doi: 10.1145/3522747.

[13] A. Sabri, C. Zouitni, H. El Medhoune, and A. Aarab, "Automatic skin lesion diagnosis and medical report generation based on image captioning," *J. Comput. Sci.*, vol. 21, no. 5, pp. 1210–1216, 2025, doi: 10.3844/jcssp.2025.1210.1216.

[14] I. Hartsock and G. Rasool, "Vision-language models for medical report generation and visual question answering: a review," *Front. Artif. Intell.*, vol. 7, Art. no. 1430984, 2024, doi: 10.3389/frai.2024.1430984.

[15] J. Pavlopoulos, V. Kougia, I. Androutsopoulos, and D. Papamichail, "Diagnostic captioning: a survey," *Knowl. Inf. Syst.*, vol. 64, no. 7, pp. 1691–1722, 2022, doi: 10.1007/s10115-022-01684-7.

[16] D.-R. Beddiar, M. Oussalah, and T. Seppänen, "Automatic captioning for medical imaging (MIC): a rapid review of literature," *Artif. Intell. Rev.*, vol. 56, no. 5, pp. 4019–4076, 2023, doi: 10.1007/s10462-022-10270-w.

[17] M. Rifki, A. Bastian, and A. Mardiana, "Development of CNN-LSTM-based image captioning dataset to enhance visual accessibility for disabilities," *JITK (Jurnal Ilmu Pengetahuan dan Teknologi Komputer)*, vol. 10, no. 4, pp. 980–992, 2025, doi: 10.33480/jitk.v10i4.6657.

[18] O. Ondeng, H. Ouma, and P. Akuon, "A review of transformer-based approaches for image captioning," *Appl. Sci.*, vol. 13, no. 19, Art. no. 11103, 2023, doi: 10.3390/app131911103.

[19] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2015, pp. 3156–3164, doi: 10.1109/CVPR.2015.7298935.

[20] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. 38th Int. Conf. Mach. Learn. (ICML)*, PMLR, vol. 139, 2021, pp. 8748–8763.

[21] Z. Zhao et al., "CLIP in medical imaging: A survey," *Med. Image Anal.*, vol. 102, Art. no. 103551, 2025, doi: 10.1016/j.media.2025.103551.

[22] S. Zhang et al., "A multimodal biomedical foundation model trained from fifteen million image–text pairs,"

NEJM AI, vol. 2, no. 1, Art. no. AIoa2400640, 2025, doi: 10.1056/AIoa2400640.

[23] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997, doi: 10.1162/neco.1997.9.8.1735.

[24] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber, "LSTM: A search space odyssey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 10, pp. 2222–2232, 2017, doi: 10.1109/TNNLS.2016.2582924.

[25] H. Fadhilah and N. P. Utama, "Systematic literature review on medical image captioning using CNN-LSTM and transformer-based models," *J. Masyarakat Informatika*, vol. 16, no. 1, pp. 32–53, 2025, doi: 10.14710/jmasif.16.1.73127.

[26] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, 1997, doi: 10.1109/78.650093.

[27] F. Hoseini and A. Y. Notash, "Image captioning using bidirectional LSTM neural network," *Discov. Artif. Intell.*, vol. 5, Art. no. 80, 2025, doi: 10.1007/s44163-025-00315-8.

[28] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," *Neural Netw.*, vol. 18, no. 5–6, pp. 602–610, 2005, doi: 10.1016/j.neunet.2005.06.042.

[29] H. E. Kim, A. Cosa-Linan, N. Santhanam, M. Jannesari, M. E. Maros, and T. Ganslandt, "Transfer learning for medical image classification: a literature review," *BMC Med. Imaging*, vol. 22, no. 1, Art. no. 69, 2022, doi: 10.1186/s12880-022-00793-7.

[30] D. M. Wonohadidjojo and L. Handayani, "Classification of Gram-stained microscopy images of bacteria cultured on multiple media using frozen and fine-tuned ResNet-50 transfer learning: A comparative study," *Teknika*, vol. 15, no. 2, pp. 357–369, 2026, doi: 10.34148/teknika.v15i2.1507.

[31] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2002, pp. 311–318, doi: 10.3115/1073083.1073135.

[32] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proc. ACL Workshop Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 2005, pp. 65–72.

[33] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*, 2004, pp. 74–81.

[34] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2015, pp. 4566–4575, doi: 10.1109/CVPR.2015.7299087.

[35] Q. Wang and A. B. Chan, "Describing like humans: On diversity in image captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 4190–4198, doi: 10.1109/CVPR.2019.00432.

[36] F. Yu et al., "Evaluating progress in automatic chest X-ray radiology report generation," *Patterns*, vol. 4, no. 9, Art. no. 100802, 2023, doi: 10.1016/j.patter.2023.100802.

[37] G. Varoquaux and V. Cheplygina, "Machine learning for medical imaging: methodological failures and recommendations for the future," *npj Digit. Med.*, vol. 5, no. 1, Art. no. 48, 2022, doi: 10.1038/s41746-022-00592-y.