Automated Chest X-Ray Captioning Using Pretrained Vision Transformer with LSTM and Multi-Head Attention

Rafy Aulia Akbar[1], Ricky Eka Putra[2], Wiyli Yustanti[3]

[1]Informatics Master Degree, Faculty of Engineering, Universitas Negeri Surabaya
[2,3]Department of Informatics, Universitas Negeri Surabaya
[1]24051905007@mhs.unesa.ac.id
[2]rickyeka@unesa.ac.id
[3]wiyliyustanti@unesa.ac.id

***Abstrak*—Radiology report generation is a complex and error-prone task, especially for radiologists with limited experience. To overcome this, this study aims to develop an automated system for generating text-based radiology reports using chest X-ray images. The proposed approach combines computer vision and natural language processing through an encoder-decoder architecture. As an encoder, a Vision Transformer (ViT) model trained on the CheXpert dataset is used to extract visual features from X-ray images after Gamma Correction is performed to improve image quality. In the decoder section, word embeddings from the report text are processed using Long Short-Term Memory (LSTM) to capture word order relationships, and enriched with Multi-Head Attention (MHA) to pay attention to important parts of the text. Visual and text features are then combined and passed to a dense layer to generate text-based radiology reports. The evaluation results show that the proposed model achieves a ROUGE-L score of 0.385, outperforming previous models. The BLEU-1 score also shows competitive results with a value of 0.427. This study shows that the use of pre-trained ViT, combined with LSTM-MHA on the decoder, provides excellent performance in capturing visual and semantic context of text, as well as improving accuracy and efficiency in radiology report automation**.

Keywords—*Vision Transformer, LSTM, Multi-Head Attention, Chest X-Ray, Medical Image Captioning.*

I. INTRODUCTION

Medical reports provide information about the patient's health condition, thus becoming a guide in determining the appropriate treatment. In addition, it is also very helpful in the clinical decision-making process and the overall management of patient care [1]. Writing a radiology report is a tiring task, as it requires expertise to accurately interpret medical images and write down findings in a structured manner in the report. This process is not only time-consuming but also prone to errors, especially for less experienced radiologists, which can lead to diagnostic inaccuracies [2], [3], [4]. High work pressure and overloaded working hours can cause fatigue, which ultimately increases the risk of misdiagnosis, which has a serious impact on the patient's health condition and potentially leads to inappropriate treatment decisions [5]. Therefore, it is necessary to utilize artificial intelligence technology that is able to provide interpretation and information from X-ray images consistently and accurately to reduce the workload experienced by radiologists. Image captioning is a complex task that combines the fields of computer vision (CV) and natural language processing (NLP), with the goal of generating descriptive text based on the visual content of an image [6]. Similar to image captioning, medical image captioning is the automated process of generating text descriptions from medical images such as X-rays to support clinical diagnosis. This field is more complex than regular image classification, as it requires a deep understanding of the relationships and interactions between elements in the image to generate relevant text [6]. This process usually uses an encoder-decoder architecture, where the encoder is responsible for extracting features from the image and the decoder generates a description that matches the input image [7]. This field usually involves two deep learning methods that include the use of Convolutional Neural Networks (CNN) for feature extraction and Recurrent Neural Network (RNN) for generating text from images [8], [9].

The main challenge in this field is the similarity of many medical images which can lead to misinterpretation in producing accurate reports. Therefore, a feature extraction model that has been trained using the chest x-ray dataset is needed to be able to distinguish each chest x-ray image. In addition, image contrast and noise can obscure important details and interfere with the accuracy of the model in recognizing medical structures. Various pre-processing techniques such as denoising, histogram equalization, and gamma correction are used to overcome this problem. Among these techniques, gamma correction is the most effective because it adjusts brightness non-linearly, increases contrast, and clarifies image details, thereby improving the quality of the resulting caption [10]. Therefore, the application of gamma correction is needed to improve image quality so that the quality of the resulting caption can be better.

In this study, before the process of extracting visual features from chest x-ray images, gamma correction was applied to improve the quality of medical images. This method is able to overcome the problem of uneven lighting, which often occurs in radiographic images. For visual feature extraction, the pretrained Vision Transformer (ViT), which has been trained using the CheXpert dataset [11], is used. After visual feature extraction, word embedding of radiology report text is also processed separately. Word embedding passes through LSTM (Long Short-Term Memory) layer to capture the sequential relationship of words in the text. The result of LSTM is then enriched with Multi-Head Attention (MHA) [12] to ensure that the model can focus on important parts of the input text. This process allows the model to understand the semantic context of the report text in more depth. Both features are then combined

before being passed to the final processing layer to generate medical report text.

In short, this study has several main contributions in the development of a deep learning-based radiology report generation system which are summarized as follows:

1) The application of pretrained Vision Transformer (ViT) which has been trained using the CheXpert dataset is carried out to extract more representative visual features.

2) The design of an efficient Encoder-Decoder architecture by applying Gamma Correction to improve image quality and using pretrained Vision Transformer as an encoder, while for text using LSTM with MHA. Both features are then combined through a fusion layer before being forwarded to the Dense layer to generate text for radiology reports.

II. Relate Works

Image captioning techniques have evolved rapidly since their early use in the medical field. Various methodologies have been proposed to improve the accuracy of the reports generated, often integrating CNN or Transformer with other models. In the early development of research in this field, CNN-based architectures as encoders for feature extraction were widely used. Research conducted by [13] utilized ResNet-152 on the encoder side with an element-wise product mechanism for image feature extraction, then combined with LSTM on the decoder. In addition, several studies also used VGG as an encoder and LSTM as a decoder, showing that although this model is quite simple, the results are still good [14].

With the development of deep learning research in this field, the Transformer architecture has begun to be widely used in the field of medical report generation. In another study, [15] used DenseNet-121 for feature extraction and Transformer as a decoder. Another study conducted by [16] used a combination of features from ResNet-50 and DenseNet-121 as a decoder and then used Meshed-Memory Transformer as a decoder. Then, (Veras Magalhães et al., 2024) proposed an approach using GPT-2 as a decoder to generate medical reports and for feature extraction using Swin Transformer. In addition, [17] introduced the use of Transformer with cross-attention mechanism to generate text descriptions and used ResNet-101 & CBAM as feature extraction.

There are studies that pay attention to the use of image enhancement to improve image quality before feature extraction, and there are also those that combine the power of CNN and Transformer as an encoder. One of the studies conducted by [10] applied Gamma Correction to improve image quality, then used DenseNet-121 (ChexNet) as an encoder for image feature extraction, while for the decoder, the study utilized BERT Embedding and Multi-Head Attention with LSTM. In addition, [18] introduced CNX-B2, a hybrid approach that uses ConvNeXt as an encoder and BioBERT as a decoder with a cross-attention mechanism to improve text generation results.

Based on previous studies, it can be seen that the use of Transformer-based encoders that have been trained using chest X-ray datasets in medical image captioning tasks is still rare. Most studies prefer to use CNN as an encoder. However, with the development of Transformer-based architectures such as ViT and Swin Transformer, efforts have begun to emerge to utilize Transformer's ability to capture long-term dependencies and global context from medical images. Even so, the application of Transformers that have been customized or retrained using radiology-specific datasets, such as CheXpert, is still limited. Therefore, this study attempts to fill this gap by using pretrained ViT that has been trained using the CheXpert dataset to improve the visual feature representation of chest X-ray images in the automated radiology report generation process.

III. Proposed Method

Figure 1 shows the architecture of the automated radiology report generation system proposed in this study. The system is designed to generate relevant text descriptions based on chest X-ray images by combining image processing, visual feature extraction, and text processing techniques. The main process in this architecture involves several important stages. First, the X-ray image goes through a gamma correction stage to enhance contrast and clarify important details in the image. After that, visual features are extracted using a pre-trained Vision Transformer (ViT), which is capable of capturing in-depth visual information from the image. In the next stage, the report text is processed using Long Short-Term Memory (LSTM) to capture the relationship of word sequences in sentences. This process is strengthened by the use of Multi-Head Attention (MHA), which allows the model to pay attention to the most relevant text sections with the extracted visual information. The combination of visual and text features is then forwarded to the dense layer to generate accurate and comprehensive text-based radiology reports. This architecture is expected to improve efficiency and accuracy in the process of automating radiology reports, providing significant support for radiologists in producing precise and fast reports.

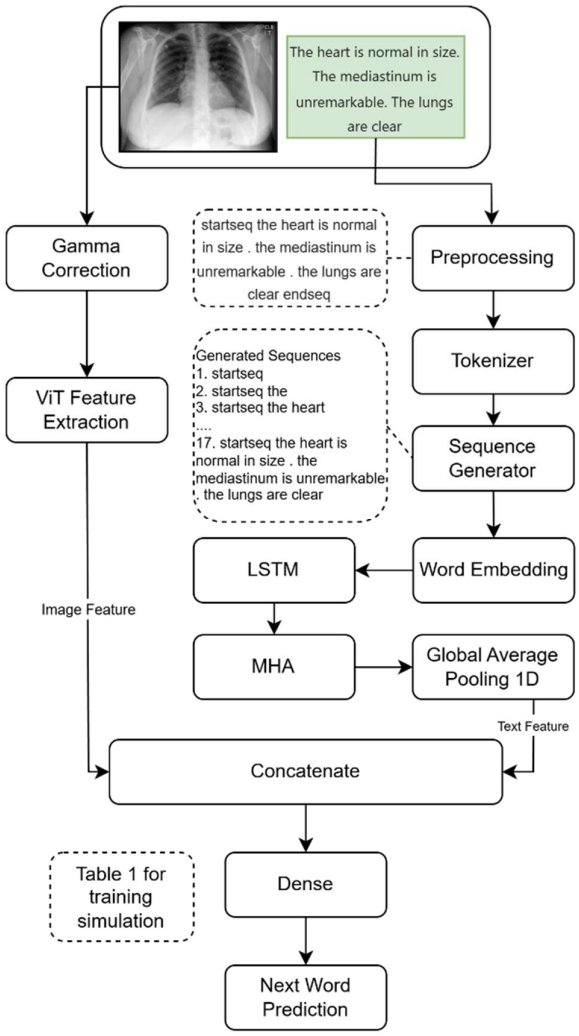


Fig. 1 Illustration of the proposed architecture.

The pre-processing steps applied to the X-ray images and radiology report texts aim to improve the quality of the data before feature extraction. As a first step on the X-ray images, gamma correction is applied to improve the contrast and brightness of the image. This technique is very important because it can improve the distribution of pixel intensity, resulting in images with more even contrast and brightness. This helps the model in capturing medical information more accurately. After the gamma correction process, the X-ray images are then given as input to the Vision Transformer (ViT), which has been trained using the CheXpert dataset. This process produces visual features that will then be used in the next process.

In parallel, the radiology report text is also processed to ensure the quality of the text data. The process begins with data cleaning, such as lowercase normalization, decontraction, removal of numbers, and non-alphabetic characters. The cleaned text is then processed using a tokenizer to convert the text into a sequence of tokens. These tokens are then converted into word embeddings using a text embedding model. The next process involves processing the text by a Long Short-Term Memory (LSTM) to capture the relationship of the word sequence in the text. The output of the LSTM is then enriched with Multi-Head Attention (MHA), which allows the model to focus on important parts of the text and understand the semantic context more deeply.

After the visual and text features are processed, the two features are combined through a concatenate layer. This concatenation allows the model to utilize visual and text information simultaneously, thereby producing more accurate and relevant text descriptions. The combined visual and text features are then fed into a dense layer to generate predictions for the next word in the radiology report. This process is carried out iteratively, with the model predicting one word at each step until the complete report is formed. Table 1 illustrates the sequence of word-by-word prediction steps, starting from the initial token "startseq" to the final token "endseq", which allows the model to generate comprehensive and precise text descriptions.

TABEL I
TRAINING SIMULATION ON AN IMAGE WITH FINDINGS

startseq the heart is normal in size . the mediastinum is unremarkable . the lungs are clear endseq	
START	
Features	**Target**
[Image Features] + [startseq]	the
[Image Features] + [startseq the]	heart
[Image Features] + [startseq the heart]	is
…	…
[Image Features] + [startseq the heart is normal in size . the mediastinum is unremarkable . the lungs are]	clear
[Image Features] + [startseq the heart is normal in size . the mediastinum is unremarkable . the lungs are clear]	enseq
END	

A. Image Enhancement

Gamma Correction is an important technique in digital image processing that functions to correct the non-linear relationship between the intensity of the input signal and the luminance produced by the display device. This is important because camera sensors and monitor screens do not have a linear response to light intensity. The main purpose of gamma correction is to align the visual appearance of the image with human perception, which has a higher sensitivity to changes in light at low luminance levels than at high luminance levels [19]. In general, the power-law method is used because of its simplicity in implementation, with the basic function $I_{out} = I_{in}^{\gamma}$, where the value of γ determines the level of correction.

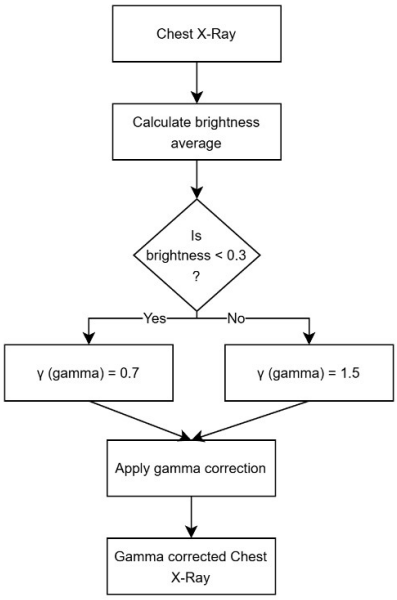


Fig. 2 Gamma Correction Flowchart

The flowchart in Figure 2 details the gamma correction process applied to chest X-ray images to improve visual quality based on brightness levels. The process begins with an input X-ray image containing important medical information. The first step is to calculate the average brightness of the image by normalizing the pixel values of the image. This normalization process aims to obtain consistent brightness values across the entire image, allowing the model to assess whether the image needs correction.

Next, if the image brightness value is below the threshold of 0.3, indicating that the image is too dark, a lower gamma value of 0.7 is applied. The use of this low gamma aims to increase the brightness of the image, so that hidden details in dark areas can be seen more clearly. Conversely, if the image is already quite bright, a higher gamma value of 1.5 is used to slightly reduce the brightness and avoid overexposure, which can obscure important details in the image.

This gamma correction process adjusts the distribution of pixel intensities in the image, improving contrast, and ensuring that relevant parts of the X-ray image, such as organ areas or body structures, are more clearly visible. Thus, the image contrast becomes more optimal, making it easier for the model to extract important visual features for further analysis. The end result of this process is a gamma-corrected chest X-ray image, with a clearer appearance and ready to be processed in the next steps in medical analysis.

In medical image processing, gamma correction is used to improve visual quality and clarify details of internal structures, which are very important in the diagnostic process. This approach is effective in reducing artifacts and increasing contrast without losing important information [20]. The use of gamma correction can significantly improve image quality by considering the characteristics of medical image classification [21].

B. Encoder

Vision Transformer (ViT) is a Transformer-based deep learning architecture designed for image processing tasks, such as image classification, and was introduced by the Google Research team in 2021 [22]. Unlike Convolutional Neural Networks (CNN) that rely on convolution filters to extract local features, ViT divides the image into small patches, such as 16×16 pixels, which are then converted into vector tokens. Each of these tokens is processed using a self-attention mechanism, which allows the model to learn the relationships between patches in the image, similar to how Transformer processes sequences of words in natural language processing (NLP). Using this mechanism, ViT can capture global patterns in the image, improving the understanding of the overall visual context. In addition, ViT utilizes a Transformer encoder consisting of self-attention and feedforward neural network layers, with positional encoding to inform the positional order of patches in the image. The main advantage of ViT lies in its ability to capture global relationships between image parts and process large-scale images with better efficiency, resulting in superior performance in various image processing tasks such as classification and object detection. This approach leverages the advantages of the Transformer architecture that was previously successfully applied to text data, opening up new opportunities in developing image processing with deeper and more complex understanding.

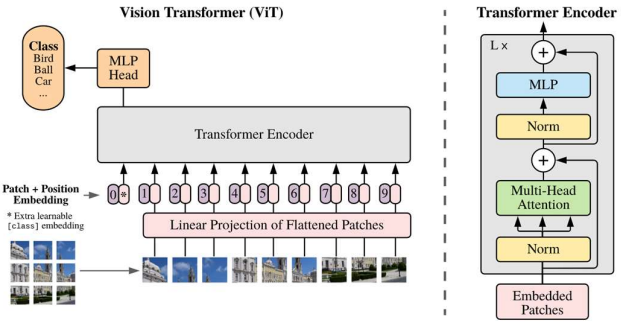


Fig. 3 Vision Transformer workflow [22]

In general, the Vision Transformer work process consists of several main stages as shown in Fig 3.

1) Image Patching: In the first stage, the input image is divided into several small patches of fixed size, for example 16×16 pixels. Each of these patches is then flattened into a one-dimensional vector and goes through a linear projection process to transform it into an initial representation ready to be processed by the model. This process allows the model to organize the image in patches, each of which stores relevant local information and can be processed independently before being combined for further analysis.

2) Positional Encoding: To preserve the spatial position information of each patch in the image, positional encoding is added to each patch representation vector. Positional encoding is essential because the Transformer model, used in Vision Transformer (ViT), has no direct understanding of the order or

position in the input data. With positional encoding, the model can understand the order and relative location of each patch in the image, thereby capturing the spatial relationships between parts of the image. This allows the model to process global information and the overall context of the image, which is key to deeper visual understanding.

3) Input Token and CLS Token: A special token called the CLS token is inserted at the beginning of the patch token sequence. This token is tasked with accumulating information from all patches to be used in the final classification process.

4) Transformer Encoder Block: Each patch token, including the CLS token, is processed through several Transformer Encoder layers.

In this study, we utilize a fine-tuned Vision Transformer (ViT) model that is publicly available on Hugging Face under the name codewithdark/vit-chest-xray. This model is a ViT-based medical image classification model trained using the CheXpert dataset, which focuses on detecting various lung disease conditions from chest X-ray images. This model has shown excellent performance during the training and validation process, achieving an accuracy of 98.46% and a low loss value, indicating its ability to classify X-ray images with a high level of accuracy. The vit-chest-xray model is used in this study as a visual feature extractor, which functions to extract important information from chest X-ray images. These extracted visual features are then used to assist the system in generating text-based radiology interpretations or reports automatically, thereby simplifying and accelerating the process of generating accurate and relevant medical reports.

C. Decoder

After the word embedding process, the vector representation of each word in the sentence will be processed by LSTM to understand the context of the word sequence as a whole. LSTM is a development of RNN designed to overcome the problem of long-term dependency with the ability to learn information in the long term through an internal memory mechanism called cell state. LSTM consists of three main gates, namely input gate, forget gate, and output gate, which dynamically regulate the flow of information. The forget gate determines which information should be forgotten, the input gate selects new information to be stored, and the output gate produces a hidden state to be forwarded to the next step. Each gate uses a combination of sigmoid and tanh activation functions to filter and transform information. This mechanism allows LSTM to retain and manipulate important information in long data sequences, making it effective for sequential data processing.

After processing using LSTM, the next step in the model is the application of Multi-Head Attention (MHA). MHA is a key component in the Transformer architecture that functions to improve the model's ability to capture contextual representations in sequential data, such as text [12]. The basic concept of MHA is the use of multiple attention mechanisms simultaneously on the input embedding. In this way, the model can pay attention to various positions in the input sequence from different perspectives, resulting in a richer and better representation compared to conventional approaches that only consider adjacent tokens. This approach allows the model to learn relationships between tokens at a greater distance in the sequence, which is important for understanding context more globally.

MHA is an extension of the self-attention or scaled dot-product attention mechanism, which calculates an attention score based on three main matrices: Query (Q), Key (K), and Value (V). This attention score is used to calculate how much attention is given to each token in the input sequence. This basic attention calculation is defined through the equation in (1), which describes how the attention value is calculated between the query and key to determine the contribution of each token to the final output. With MHA, the model can parallelly capture contextual information from different parts of the input, enhancing the model's ability to understand data sequences more comprehensively and effectively.

$$Attention\ (Q,K,V) = softmax\ \left(\frac{Q\ K^T}{\sqrt{d_k}}\right)V \quad (1)$$

Where d_k is the dimension of Key. This mechanism allows each element in the input sequence to give attention weight to other elements in the sequence. MHA strengthens this ability by dividing the attention process into several heads that work in parallel. Each head uses different projection weights to form different sets of Q, K, and V. Each head calculates self-attention independently, then the results of all heads are combined (concate) and projected back with the final weight matrix Wo. Mathematically, MHA can be defined in Equation 2 and 3.

$$MultiHead(Q,K,V) = [head_1, head_2, \dots, head_h]W^O \quad (2)$$

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \quad (3)$$

Where W^O is the final weight matrix which functions to change the combined results of all heads into a form or dimension that suits the model's needs.

IV. EXPERIMENTAL

The proposed methodology for medical report generation is done using IU X-Ray dataset and all the implementation details are described in detail in this section.

A. Implementation Details

The proposed methodology is implemented using the Python programming language by utilizing the Keras framework (with TensorFlow backend) and PyTorch for the integration of the Hugging Face library. For the encoder, the Vision Transformer (ViT) model that has been trained using the CheXpert dataset [11] is used. The training process of this model is carried out on hardware consisting of an Intel Core i9-13900H, 32 GB of RAM, and an NVIDIA RTX 4060 GPU with 8 GB of memory, with a total training time of 25 epochs using a batch size of 1. This model applies the LeakyReLU activation function to avoid the “neuron death” phenomenon [23] and uses the Categorical Crossentropy loss function. During training, the

ReduceLROnPlateau technique is applied with a learning rate reduction factor of 0.5 and a patience of 4, followed by the use of a dropout layer with a rate of 0.5 to reduce the risk of overfitting [23]. Each epoch of the model is saved and tested against all evaluation metrics, so that the best epoch can be selected based on the results of the evaluation of these metrics.

The training process was carried out for 25 epochs with a batch size of 1, resulting in a total training time of 5 hours 7 minutes 44.91 seconds, a relatively short time considering the complexity of the model used. This training speed is supported by the application of image augmentation, image enhancement, and feature extraction which are carried out off-line before the training process begins. The average time required for each epoch is around 738.58 seconds (around 12 minutes 18.58 seconds). However, at the beginning of training (the first and second epochs), the model showed low scores or even failed to form meaningful sentences, which caused the duration per epoch to exceed 2 hours. Several other epochs were also recorded to exceed 25 minutes. After that, the duration per epoch began to stabilize in the range of 16 to 25 minutes per epoch. The model evaluation process, which involves word prediction and the formation of complete sentences, requires an additional time of around 12 hours 17 minutes. Thus, the total time required for model training and evaluation reached 17 hours 24 minutes, reflecting a fairly efficient time for this high-complexity process.

B. Evaluation Metrics

Based on many previous studies, to evaluate the ability of the model in generating radiology reports, we use several evaluation metrics commonly used in the fields of image captioning and medical report generation, namely BLEU [24] and ROUGE-L [25]. The BLEU metric, although originally developed for language translation tasks, has been widely adopted for evaluating the quality of texts produced by research in this field. Meanwhile, ROUGE-L is used to measure how similar the text predicted by the model is to the reference text based on the same word sequence and sequence between the two texts.

C. Dataset

The data used in this study is the IU X-ray Dataset [26], which consists of chest X-ray images paired with their associated diagnostic reports. This dataset was collected by Indiana University (IU) and contains 7,470 pairs of X-ray images and 3,955 reports in XML format, accompanied by problem tags that describe the medical conditions identified in the images. The images in this dataset are chest X-rays in a standard medical imaging format, while the reports are diagnostic descriptions written by radiologists in text form. The images in this dataset vary in size, but have generally been adjusted for analysis purposes. Preprocessing has been carried out to ensure consistency of format and data quality, including adjusting the image size to suit the analysis needs. This dataset comes from a trusted source and has gone through various stages of validation to ensure the accuracy and reliability of the data used in the study. Figure 4 shows a sample example from the IU X-ray Dataset, depicting pairs of X-ray images and their relevant diagnostic reports.

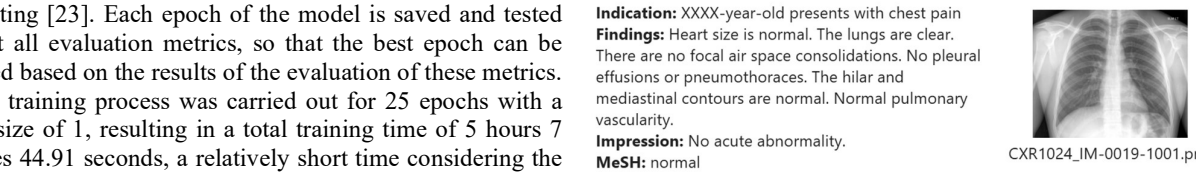


Fig. 4 Samples from the IU dataset consisting of Indication, Findings, Impression, and MeSH

Each report may not have images and some are associated with two to five images. The report includes impressions and findings as reports. In this study, the findings section is used as the target text for the generation process following previous research. In addition, the data is divided into two main parts: 90% for training data and 10% for test data. On the training data, image augmentation is carried out using a rotation technique (maximum 5 degrees) with a probability of 50%, as well as random brightness and contrast modifications with a limit of ±0.05 and the same probability. This augmentation is applied to increase data variation in images with low tag frequencies. Meanwhile, text pre-processing includes normalization to lowercase, decontraction, removal of numbers, and non-alphabetic characters to improve text quality before the modeling stage.

V. Result

In this section, we present the results of the study, including visualization of images processed using the adaptive gamma correction method, as well as a quantitative evaluation that measures the performance of the model in generating automated radiology reports. This evaluation is carried out by comparing the BLEU score and ROUGE-L values between the reports generated with and without the application of adaptive gamma correction. This process allows us to measure the extent to which the improvement in image quality resulting from adaptive gamma correction contributes to improving the accuracy of the model in generating more relevant and precise medical descriptions. In addition, we also compare the results obtained with previous studies using similar methods, to assess the effectiveness of the proposed approach. This comparison provides deeper insight into the advantages or potential disadvantages of the adaptive gamma correction method in the context of automated radiology reports, as well as comparing its contribution to the performance of the model in generating more accurate reports, especially on images with low quality or less clear details. Thus, this section not only provides an overview of the effectiveness of the proposed technique, but also places it in the context of the existing literature, providing a more comprehensive perspective on the progress made in this field.

A. Image Enhancement

Before the visual feature extraction process is carried out, the quality of the CXR image is enhanced using the Gamma Correction method. The main purpose of this stage is to

increase the contrast and clarify important details in the image, making it easier for the model to capture more accurate medical information. Gamma Correction functions to adjust the brightness of pixels based on the gamma value, which in turn improves the contrast of the image, especially in areas with low lighting or contrast. Fig 5 shows a comparison between a chest X-ray image in normal conditions (unprocessed) and after Gamma Correction is applied. From this comparison, significant differences can be observed in both images, where the image that has been applied with Gamma Correction shows a clear increase in contrast, especially in certain parts that were previously unclear or too dark. This process allows previously hidden or less visible details to be more clearly visible, which is very important in the context of medical diagnosis. The benefits of using Gamma Correction as a contrast enhancement method for medical images are very visible in improving image quality, which in turn helps the model to perform a more accurate and efficient analysis of the medical conditions present in the X-ray image.

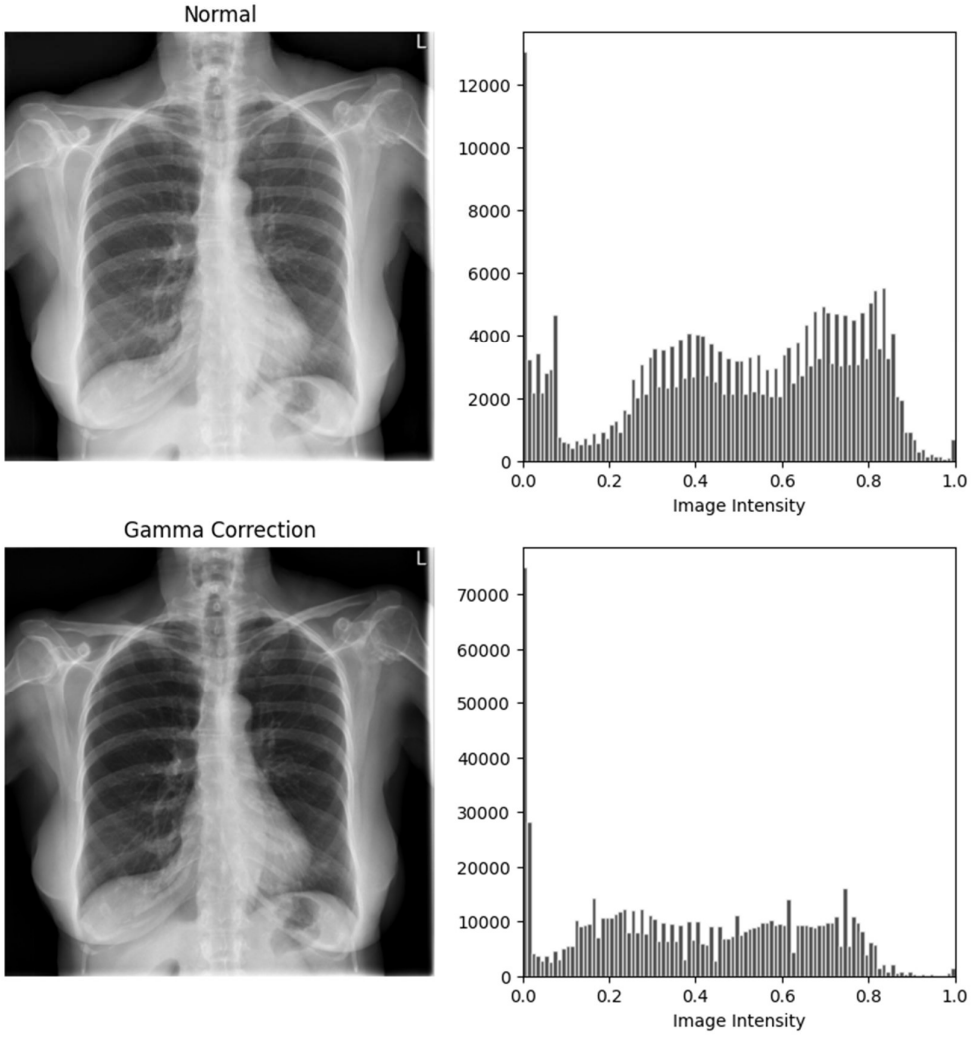


Fig. 5 Comparison results between normal images and gamma correction

In the upper left part of Figure 5, there is a chest X-ray image in normal condition. This image has relatively low contrast. The histogram presented on the right shows the pixel intensity distribution that tends to be skewed towards the dark side, with the majority of the intensity falling in a narrow range of values. This indicates that the original image has uneven color variations, so some important details may be overlooked or difficult to extract by the model. After Gamma Correction is applied, the chest X-ray image (bottom left part of Figure 5) experiences a significant increase in contrast. Anatomical details such as the lungs, ribs, and other organs become clearer. This helps improve the visibility of important features in the image, which is very important for the radiological interpretation process. The histogram in the lower right part of Figure 5 also shows a striking change: the pixel intensity distribution becomes more even and wider, with a wider spread of intensity values. This shows that Gamma Correction successfully optimizes the pixel intensity distribution, thereby

improving the visual quality and information available in the image.

B. Model Outputs

In this section, we present the results of the model performance evaluation in generating automated radiology reports using standard evaluation metrics, namely BLEU and ROUGE-L. The purpose of this evaluation is to measure the extent to which the reports generated by the model can match or replicate the reference reports prepared by radiologists. The BLEU (Bilingual Evaluation Understudy) metric measures the n-gram similarity between the text generated by the model and the reference text. This metric provides an overview of the model's ability to replicate relevant words, phrases, and structures from the original report, as well as how well the model captures important elements in the reference text. As an n-gram-based metric, BLEU focuses more on surface similarity and is often used in machine translation systems. On the other hand, ROUGE-L (Recall-Oriented Understudy for Gisting Evaluation) assesses similarity based on the longest common subsequence between the model-generated text and the reference. This metric is more sensitive to sentence structure and overall context, thus providing a more comprehensive assessment of the quality of the match between the generated report and the reference. ROUGE-L pays more attention to the similarity of meaning and relationships between sentences in the report, which is very important in the context of radiology report generation that requires precise contextual understanding. These two metrics, BLEU and ROUGE-L, provide comprehensive insights into the model's ability to generate reports that are not only technically accurate but also clinically relevant, reflecting the patient's condition in a way that is close to a real radiologist's report.

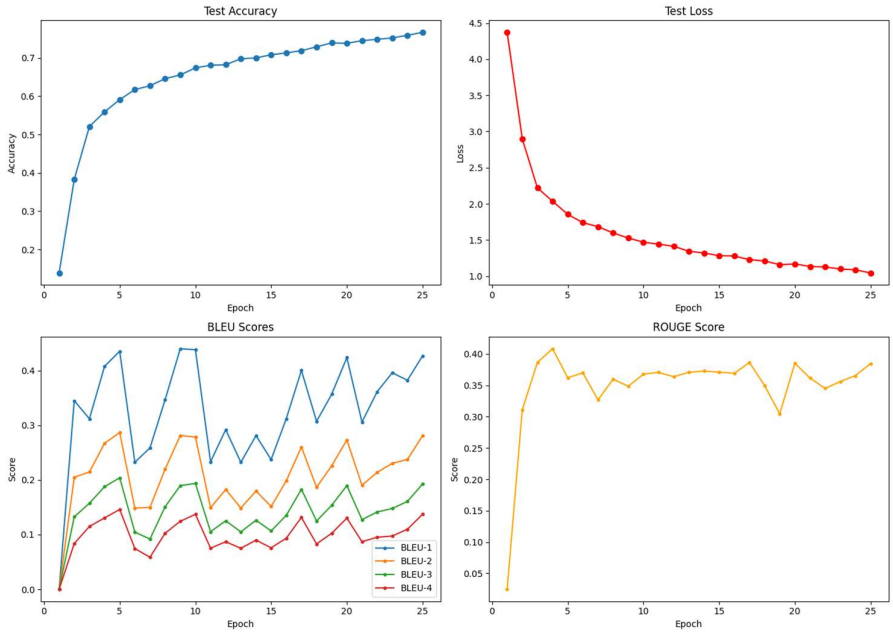


Fig. 6 Accuracy, loss, and score during training

The first graph in Figure 6, Test Accuracy, shows a significant increase in model accuracy over time. In the beginning (above epoch 0), the accuracy value is very low, but increases sharply to above 70% at epoch 25. The second graph, Test Loss, shows that at the beginning of training, the loss value is very high, but drops drastically to stabilize at around 1.0 at the last epoch. The BLEU Scores graph shows a consistent increase for all BLEU variables (BLEU-1 to BLEU-4) although there are some fluctuations, with BLEU-1 having the highest value and BLEU-4 lower due to the complexity of the long text. Finally, the ROUGE Score graph also shows an increase in ROUGE-L to reach 0.40 at certain epochs, indicating that the prediction results are getting closer to the reference report.

TABEL III
COMPARATIVE RESULTS WITH PREVIOUS RESEARCH

MODEL	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L
Full-ARL [2]				0.125	0.262
ResNet 50 + DenseNet 121 & MMF [16]	0.348	0.218	0.15	0.106	0.237
CNX-B2 [18]	**0.479**	**0.363**	**0.261**	**0.173**	0.354
XraySwinGen [27]	0.377	0.239	0.168	0.124	0.3
Ours	0.427	0.281	0.193	0.137	**0.385**

In Table 2, the BLEU-1 metric shows that the CNX-B2 model [18] obtained the highest score of 0.479, while the proposed model scored 0.427, which although slightly lower, is still in the competitive range compared to other tested models. However, when viewed overall from the BLEU-1 to BLEU-4

metrics, the proposed model shows superior performance. This can be seen in the comparison between the proposed model and the Full-ARL [2], MMF [16], and XraySwinGen [27] models. For example, in the BLEU-4 metric, the proposed model obtained a score of 0.137, higher than the Full-ARL model which obtained a score of 0.125 and XraySwinGen which only reached 0.124. In addition, in the ROUGE-L metric which measures the level of agreement between the predicted text and the reference text based on the longest subsequence, the proposed model also shows outstanding performance. With a ROUGE-L score of 0.385, the model outperforms all previous models, indicating its ability to produce more relevant and accurate results. Overall, despite not ranking top on all metrics, the proposed model performs very competitively, even better in some important metrics compared to existing leading models.

Fig 7 shows a clear comparison between the radiology reports generated by the model (Prediction) and the reference report (Ground Truth) based on three chest X-ray (CXR) images. This comparison provides an overview of the extent to which the model is able to generate relevant, accurate, and appropriate text descriptions of the analyzed medical images. Each example of the predicted report is compared with the ground truth report, accompanied by an assessment of the BLEU-1 to BLEU-4 evaluation metrics as well as ROUGE-L, which provide a quantitative overview of the model's performance in producing medical descriptive texts. Overall, the results shown in Figure 6 indicate that the proposed model successfully generates automatic reports that are quite close to the reference report, with relatively minor differences, such as additional phrases that do not significantly affect the understanding of the report content. The generated predicted reports remain relevant and include important medical information that should be in the radiology report, according to the patient's condition being analyzed. This success is supported by the fairly high results of the BLEU and ROUGE-L evaluation metrics, indicating that the proposed approach is effective in combining visual features of X-ray images with textual information to generate accurate and comprehensive medical descriptions. However, a deeper analysis reveals that the model still faces challenges in generating accurate reports for anomalous cases or abnormal medical conditions. This is largely due to the high data imbalance between normal and abnormal findings in the training dataset, where normal findings are more dominant. As a result, the model tends to be better trained to predict common and easily recognizable findings, such as normal conditions, but is less effective in identifying or describing rare or complex conditions that are abnormal. The model's inability to handle such cases affects the model's generalizability to more varied and diverse data, which is an important area for further model improvement and development.

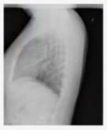
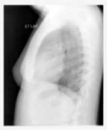
CXR	Ground Truth	Prediction	Scores
	the heart is normal in size . the mediastinum is unremarkable . the lungs are clear . no evidence of active tuberculosis . no acute disease	the heart is normal in size . the mediastinum is unremarkable . the lungs are clear . no focal airspace consolidation . no pleural effusion . no acute cardiopulmonary abnormality	**bleu1**: 0.7188 **bleu2**: 0.6810 **bleu3**: 0.6556 **bleu4**: 0.6259 **rouge**: 0.7667
	the heart is normal in size . the mediastinum is unremarkable . the lungs are clear . no acute disease	the heart is normal in size . the mediastinum is unremarkable . the lungs are clear . no focal airspace consolidation . no pleural effusion . no acute pulmonary disease	**bleu1**: 0.6875 **bleu2**: 0.6660 **bleu3**: 0.6461 **bleu4**: 0.6190 **rouge**: 0.8148
	the heart is normal in size . the mediastinum is unremarkable . the lungs are clear . no acute disease	the heart is normal in size . the mediastinum is unremarkable . the lungs are clear . no focal airspace consolidation . no pleural effusion . no acute pulmonary abnormality	**bleu1**: 0.6562 **bleu2**: 0.6342 **bleu3**: 0.6256 **bleu4**: 0.6041 **rouge**: 0.7778

Fig. 7 Example between prediction results and ground truth

VI. CONCLUSION

This study proposes an automated system for radiology report generation based on visual feature extraction using Vision Transformer (ViT) and a text decoder consisting of Long Short-Term Memory (LSTM) and Multi-Head Attention (MHA). This system aims to improve the efficiency and accuracy in generating clinically relevant and linguistically coherent radiology reports, by leveraging state-of-the-art technologies in medical image processing and natural language processing.

Experimental results show that the proposed model is able to outperform other models in terms of ROUGE-L score and achieve competitive results in BLEU score, indicating that this system is effective in generating reports that are not only accurate but also linguistically coherent. The application of Gamma Correction in the image preprocessing stage is also proven to improve the quality of chest X-ray images, allowing for more accurate feature extraction and supporting better classification and reporting processes.

However, this study also has some limitations that need to be improved in the future. The model used was trained with a limited dataset, which limits its generalization ability,

especially in detecting rare findings. In addition, the imbalance between normal and abnormal data and the use of fixed Gamma Correction parameters are also challenges. Therefore, further research will focus on the use of larger and more balanced datasets, as well as the exploration of adaptive image enhancement and data augmentation techniques to overcome the imbalance problem and improve image quality.

VII. REFERENSI

[1] K. L. Gormly, "Improving radiology reporting locally and globally: who, how, and why?," *British Journal of Radiology*, vol. 98, no. 1167, pp. 330–335, Mar. 2025, doi: 10.1093/bjr/tqae253.

[2] J. Yuan, H. Liao, R. Luo, and J. Luo, "Automatic Radiology Report Generation Based on Multi-view Image Fusion and Medical Concept Enrichment," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2019. doi: 10.1007/978-3-030-32226-7_80.

[3] P. Divya, Y. Sravani, C. Vishnu, C. K. Mohan, and Y. W. Chen, "Memory Guided Transformer With Spatio-Semantic Visual Extractor for Medical Report Generation," *IEEE J Biomed Health Inform*, vol. 28, no. 5, 2024, doi: 10.1109/JBHI.2024.3371894.

[4] Z. Chen, Y. Song, T. H. Chang, and X. Wan, "Generating radiology reports via memory-driven transformer," in *EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, 2020. doi: 10.18653/v1/2020.emnlp-main.112.

[5] Y. C. Peng, W. J. Lee, Y. C. Chang, W. P. Chan, and S. J. Chen, "Radiologist burnout: Trends in medical imaging utilization under the national health insurance system with the universal code bundling strategy in an academic tertiary medical centre," *Eur J Radiol*, vol. 157, 2022, doi: 10.1016/j.ejrad.2022.110596.

[6] S. Elbedwehy, T. Medhat, T. Hamza, and M. F. Alrahmawy, "Enhanced descriptive captioning model for histopathological patches," *Multimed Tools Appl*, vol. 83, no. 12, 2024, doi: 10.1007/s11042-023-15884-y.

[7] Z. Song and X. Zhou, "EXPLORING EXPLICIT AND IMPLICIT VISUAL RELATIONSHIPS FOR IMAGE CAPTIONING," in *Proceedings - IEEE International Conference on Multimedia and Expo*, 2021. doi: 10.1109/ICME51207.2021.9428310.

[8] M. Liu, H. Hu, L. Li, Y. Yu, and W. Guan, "Chinese Image Caption Generation via Visual Attention and Topic Modeling," *IEEE Trans Cybern*, vol. 52, no. 2, 2022, doi: 10.1109/TCYB.2020.2997034.

[9] H. Chen, G. Ding, Z. Lin, Y. Guo, C. Shan, and J. Han, "Image Captioning with Memorized Knowledge," *Cognit Comput*, vol. 13, no. 4, 2021, doi: 10.1007/s12559-019-09656-w.

[10] H. Tsaniya, C. Fatichah, and N. Suciati, "Automatic Radiology Report Generator Using Transformer With Contrast-Based Image Enhancement," *IEEE Access*, vol. 12, 2024, doi: 10.1109/ACCESS.2024.3364373.

[11] J. Irvin *et al.*, "CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison." [Online]. Available: www.aaai.org

[12] A. Vaswani *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017.

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

[14] Z. Babar, T. van Laarhoven, and E. Marchiori, "Encoder-decoder models for chest X-ray report generation perform no better than unconditioned baselines," *PLoS One*, vol. 16, no. 11 November, 2021, doi: 10.1371/journal.pone.0259639.

[15] S. Yan, W. K. Cheung, K. Chiu, T. M. Tong, K. C. Cheung, and S. See, "Attributed Abnormality Graph Embedding for Clinically Accurate X-Ray Report Generation," *IEEE Trans Med Imaging*, vol. 42, no. 8, 2023, doi: 10.1109/TMI.2023.3245608.

[16] E. Vendrow and E. Schonfeld, "Understanding transfer learning for chest radiograph clinical report generation with modified transformer architectures," *Heliyon*, vol. 9, no. 7, 2023, doi: 10.1016/j.heliyon.2023.e17968.

[17] J. Zhao *et al.*, "Automated Chest X-Ray Diagnosis Report Generation with Cross-Attention Mechanism," *Applied Sciences (Switzerland)*, vol. 15, no. 1, Jan. 2025, doi: 10.3390/app15010343.

[18] F. F. Alqahtani, M. M. Mohsan, K. Alshamrani, J. Zeb, S. Alhamami, and D. Alqarni, "CNX-B2: A Novel CNN-Transformer Approach For Chest X-Ray Medical Report Generation," *IEEE Access*, vol. 12, 2024, doi: 10.1109/ACCESS.2024.3367360.

[19] S. Soni, P. Singh, and A. A. Waoo, "REVIEW OF GAMMA CORRECTION TECHNIQUES IN DIGITAL IMAGING," *ShodhKosh: Journal of Visual and Performing Arts*, vol. 5, no. 5, May 2024, doi: 10.29121/shodhkosh.v5.i5.2024.1902.

[20] U. K. Acharya and S. Kumar, "Directed searching optimized texture based adaptive gamma correction (DSOTAGC) technique for medical image enhancement," *Multimed Tools Appl*, vol. 83, no. 3, pp. 6943–6962, Jan. 2024, doi: 10.1007/s11042-023-15953-2.

[21] F. Kallel and A. Ben Hamida, "A New Adaptive Gamma Correction Based Algorithm Using DWT-SVD for Non-Contrast CT Image Enhancement," *IEEE Trans Nanobioscience*, vol. 16, no. 8, pp. 666–675, 2017, doi: 10.1109/TNB.2017.2771350.

[22] A. Dosovitskiy *et al.*, "AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE," in *ICLR 2021 - 9th International Conference on Learning Representations*, 2021.

[23] W. Liu, J. Luo, Y. Yang, W. Wang, J. Deng, and L. Yu, "Automatic lung segmentation in chest X-ray images using improved U-Net," *Sci Rep*, vol. 12, no. 1, 2022, doi: 10.1038/s41598-022-12743-y.

[24] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a Method for Automatic Evaluation of Machine Translation."

[25] C.-Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries."

[26] D. Demner-Fushman *et al.*, "Preparing a collection of radiology examinations for distribution and retrieval," *Journal of the American Medical Informatics Association*, vol. 23, no. 2, 2016, doi: 10.1093/jamia/ocv080.

[27] G. Veras Magalhães, R. L. de S. Santos, L. H. S. Vogado, A. Cardoso de Paiva, and P. de Alcântara dos Santos Neto, "XRaySwinGen: Automatic medical reporting for X-ray exams with multimodal model," *Heliyon*, vol. 10, no. 7, 2024, doi: 10.1016/j.heliyon.2024.e27516.

.