Indonesian Journal of Electronics, Electromedical Engineering, and Medical Informatics e-ISSN: 2656-8624
Homepage: https://ijeeemi.org/; Vol. 7, No. 4, pp. 654-672, November 2025

RESEARCH PAPER **OPEN ACCES**

CoAtNet for Chest X-Ray Report Generation with Bi-LSTM and Multi-Head Attention

Rafy Aulia Akbar, Ricky Eka Putra, and Wiyli Yustanti
Department of Informatics, Universitas Negeri Surabaya, Surabaya, East Java, Indonesia

Abstract

In clinical environments, chest X-Ray (CXR) represents one of the most prevalent diagnostic instruments, particularly facilitating diagnostic procedures through medical reports. However, manual report preparation is time-consuming, highly dependent on the expertise of radiologists, and carries the risk of errors due to heavy workloads and limited expert staff. Therefore, an automated system based on artificial intelligence is needed to ease the workload of radiologists while improving consistency. This study aims to develop an automated medical report generation system with balanced data distribution, a reliable encoder, and bidirectional contextual understanding. The main contributions of this study include the implementation of an undersampling strategy based on majority captions, followed by oversampling of minority labels while maintaining the proportion of labels with higher frequencies; the use of a Bi-LSTM with Multi-Head Attention (MHA) to strengthen contextual understanding in text; and the use of CoAtNet as a visual encoder that combines the strengths of CNNs and Transformers. The methodology incorporates image preprocessing via gamma correction for contrast enhancement, data selection, balancing through combined undersampling and oversampling, and CoAtNet implementation as an encoder paired with Bi-LSTM and MHA as the decoder. Experimental evaluation employed the IU X-ray dataset, with performance assessed using BLEU and ROUGE-L metrics. The results revealed that the CoAtNet configuration with Bi-LSTM and MHA, coupled with the undersampling–oversampling strategy, delivered superior performance, evidenced by a cumulative score of 1.642, with BLEU-1 to BLEU-4 and ROUGE-L achieving 0.480, 0.329, 0.245, 0.183, and 0.405, respectively. These findings demonstrate that the combination of data balancing strategies with CoAtNet and Bi-LSTM can produce more accurate automated medical reports and reduce bias toward majority labels.

Paper History

Received August 22, 2025
Revised October 10, 2025
Accepted October 20, 2025
Published November 30, 2025

Keywords

Medical Image Captioning;
Chest X-Ray Report Generation;
Gamma Correction;
CoAtNet;
Bi-LSTM;
Multi-Head Attention;
Data Imbalance Handling

Author Email

24051905007@mhs.unesa.ac.id
rickyeka@unesa.ac.id
wiyliyustanti@unesa.ac.id

I. Introduction

CXR stands as one of the predominant diagnostic instruments in clinical settings [1], [2]. CXR reports facilitate vital communication between radiologists and clinicians, delivering key information for diagnosis, treatment planning, and patient care [3], [4]. In practice, the process of manually writing medical reports from chest radiographic images is time-consuming, relies on radiologists' expertise that may be unavailable in remote or resource-limited areas, and can lead to inter-author variability [5], [6], [7]. This situation is exacerbated by the increasing number of patients and the limited availability of experts, particularly in remote regions [8]. Misdiagnosis resulting from inaccurate reports can have serious consequences for patient health and clinical decision-making [9], [10]. Consequently, advancements in artificial intelligence (AI) technologies have become imperative to develop automated solutions for CXR analysis, reduce the workload of radiologists, and enhance the reliability and precision of medical documentation. One of the most relevant AI approaches is image captioning, a research field that aims to generate automatic textual descriptions of images [11].

Image captioning represents an interdisciplinary field that merges computer vision and natural language processing within a multimodal learning framework, focusing on producing textual descriptions from visual inputs [12]. This task is more complex than ordinary image classification because it requires a deep understanding of the spatial and semantic relationships between objects in the image to generate relevant narratives [13], [14]. In general, image captioning frameworks employ an encoder–decoder architecture, where the encoder often implemented as a CNN extracts visual features from images, while the decoder commonly utilizing RNN or LSTM generates descriptive text sequences [15], [16]. With technological advancements, this approach has been adapted to the medical domain as medical image captioning or medical report generation, which automatically produces radiology reports based on medical images such as CXR.

Several previous studies have developed captioning models for CXR images; however, they still exhibit performance limitations and have not fully addressed fundamental challenges. The research in [17] implemented a ResNet-152 encoder paired with an LSTM decoder architecture and, when evaluated on the IU X-ray

Corresponding author: Ricky Eka Putra, rickyeka@unesa.ac.id, Department of Informatics, State University of Surabaya, Jl. Ketintang, Ketintang, Kec. Gayungan, Surabaya, East Java 60231, Indonesia.
Digital Object Idenfier (DOI): https://doi.org/10.35882/ijeeemi.v7i4.271

dataset, reported BLEU-4 (B-4) and ROUGE-L (R-L) scores of 0.101 and 0.293, respectively. This indicates that although the element-wise product method outperformed global average pooling, the model still failed to accurately describe abnormal findings due to data imbalance. Meanwhile, the research in [18], which applied VGG-19 and LSTM on the same dataset, achieved a B-4 of 0.115 and an R-L of 0.276, without showing significant improvement over the baseline or incorporating any enhancement strategies to address these weaknesses.

Previous investigations have sought to enhance the quality of automated medical report generation through advanced architectural designs and attention-based techniques, yet these approaches continue to exhibit fundamental limitations. The research in [19], which proposed an ARL framework based on ResNet-152 and LSTM with multi-level attention, achieved only a B-4 of 0.125 and an R-L of 0.262, indicating that the model remained susceptible to data bias and was unable to fully address the risk of hallucinations. Meanwhile, the research in [20], which implemented ResNet-50 and dual-LSTM with Cross-View Attention Module (CVAM), recorded improved performance with a B-4 of 0.152 and an R-L of 0.385 but remained constrained by the narrow vocabulary of the IU X-ray dataset, thereby failing to generate new medical terms that did not appear during training. The research in [21], which combined DenseNet-121 with an abnormality graph representation and a Two-Level LSTM or Transformer decoder, successfully improved clinical accuracy with an R-L of 0.374 but still struggled to distinguish similar visual abnormalities due to reliance on ambiguous and inconsistent annotations.

Recent studies have begun integrating image enhancement techniques, yet they still suffer from fundamental weaknesses in contextual understanding and generalization. The research in [22], which used gamma correction to enhance images, CheXNet as an encoder, and BERT and LSTM with Multi-Head Attention, reported BLEU metric results of 0.363 for B-1, 0.371 for B-2, 0.388 for B-3, and 0.412 for B-4, respectively an anomaly likely caused by misreporting, as in practice, BLEU scores should decrease as the n-gram order increases. Despite these numerical results, the model remained limited by the use of a unidirectional LSTM that failed to optimally capture the global context of the text [23], [24]. Meanwhile, the research in [25], which leveraged the Swin Transformer and BERT with text augmentation, achieved only a B-4 of 0.151 and an R-L of 0.343, indicating that although text augmentation helps mitigate data imbalance, the model still struggled to generate complete and diverse descriptions.

Further studies began adopting Transformer architectures and large language models but faced computational complexity constraints. The research in [26], which combined ConvNeXt and BioBERT with Cross Attention, achieved competitive performance with a B-4 score of 0.173 and an R-L of 0.354. However, this model had high complexity (224.31 million parameters, 222 GFLOPS), making it difficult to implement on resource-constrained systems. Meanwhile, the research in [27], which used Swin Transformer and GPT-2, achieved only a B-4 of 0.124 and an R-L of 0.300 and was at high risk of producing hallucinations due to the limited representation and length of descriptions in the training data. Finally, the approach in [28], which integrated ResNet-101 with CBAM and Transformer, achieved a B-4 of 0.152 and an R-L of 0.364 but was less capable of handling complex medical cases.

Despite various approaches being developed, three fundamental challenges remain incompletely addressed. First, the dominance of the normal label in datasets causes models to become biased and less sensitive to abnormal findings [17], [19], [25]. Second, some conventional visual encoders fail to extract fine-grained features necessary to distinguish similar radiological abnormalities [17], [21]. Third, unidirectional LSTM architectures limit contextual text understanding, as they cannot capture bidirectional semantic relationships [17], [18], [20], [22]. Therefore, this study proposes an approach to address all three issues through data balancing, a spatial-detail-sensitive hybrid visual encoder, and a bidirectional contextual decoder architecture.

In response to these challenges, this study introduces three key innovations through the implementation of an automated medical report generation framework. First, to mitigate data bias, we introduce a methodology that combines undersampling of prevalent captions with oversampling of underrepresented labels while preserving the proportion of majority labels. Notably, we avoid synthetic oversampling methods such as SMOTE, as they generate feature vectors through numerical interpolation without corresponding CXR images, thereby breaking the essential image–text alignment required in medical report generation. Second, as a contribution to visual feature extraction, we adopt CoAtNet [29], a hybrid CNN-Transformer architecture proven superior in disease detection from CXR images [30], as the visual encoder. Third, we replace the unidirectional LSTM with a Bi-LSTM, followed by an MHA layer. Implementing Bi-LSTM enhances the model's capacity to capture bidirectional semantic context and word relationships more accurately. For example, in conventional image captioning tasks, the research in [24] demonstrated a 9.7% performance improvement compared with using an LSTM. Moreover, integrating Bi-LSTM with an attention mechanism has consistently been shown to enhance performance across various evaluation metrics [31].

This manuscript is structured into five sections. Section II elaborates on the dataset, preprocessing techniques for images and text, data imbalance mitigation, and the implemented architecture. Section III presents the experimental results. Section IV provides a comprehensive analysis of the results and discusses their implications. Section V summarizes the principal research contributions and suggests potential directions for future work.

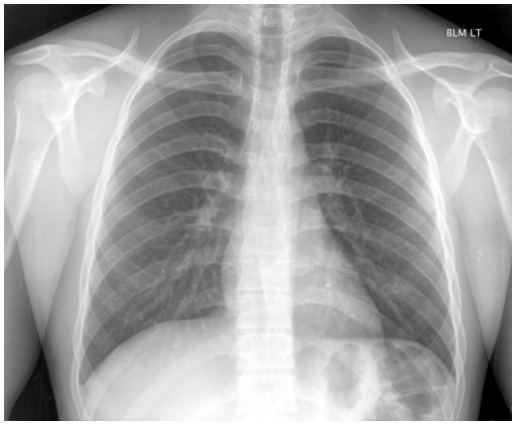
II. MATERIALS AND METHOD

A. Dataset

We utilized the IU X-ray dataset provided by [32], which contains 3,955 medical reports and 7,470 DICOM

Corresponding author: Ricky Eka Putra, rickyeka@unesa.ac.id, Department of Informatics, State University of Surabaya, Jl. Ketintang, Ketintang, Kec. Gayungan, Surabaya, East Java 60231, Indonesia.
Digital Object Idenfier (DOI): https://doi.org/10.35882/ijeeemi.v7i4.271

images. This dataset classifies images based on radiographic projection into three categories: 3,820 images with posteroanterior (PA) projection, 3,646 images with lateral projection, and four images with unclear or unidentifiable projection types. Each image has a varying resolution, and all images have been normalized to meet the requirements of computational analysis.



CXR1024_IM-0019-1001

Comparison	None
Indication	XXXX-year-old presents with chest pain
Findings	The heart size is normal. The lungs are clear. There is no focal airspace consolidation. No pleural effusion or pneumothorax is seen. The hilar and mediastinal contours are normal. Pulmonary vascularity is normal.
Impression	No acute abnormality.
MeSH	normal
Problems	normal
Tags	normal

Fig. 1. Example of an IU X-Ray dataset consisting of CXR image, Comparison, Indication, Findings, Impression, MeSH, Problems, and Tags.

Table 1. Number of image-report pairs in each split of the IU X-ray Dataset.

Split	Quantity
Train	4128
Validation	516
Test	516

As an illustration, Fig. 1 presents an example of a sample visualization from the dataset. Each report in this dataset contains a radiology report that can be associated with zero to five images, and the report consists of two main components: impressions and findings. This dataset is the most widely used dataset in research within this field [33] and is publicly available at https://openi.nlm.nih.gov and on the Kaggle platform at https://www.kaggle.com/datasets/raddar/chest-xrays-indiana-university. In this experiment, we used the findings section as the target because this approach is commonly adopted in previous research [19], [20], [25], [26]. The dataset was divided into three subsets: training, validation, and testing, with the specific distribution detailed in Table 1.

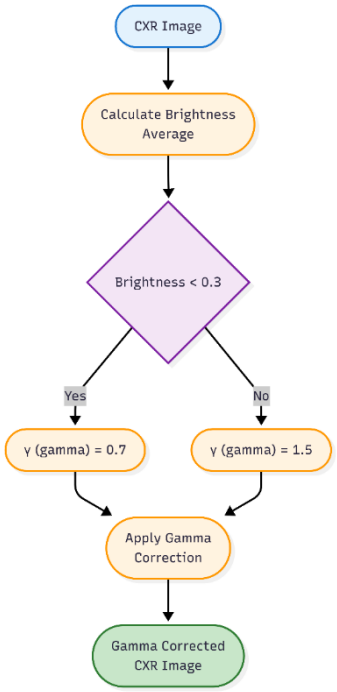


Fig. 2. Flowchart of gamma correction technique implementation for contrast enhancement of CXR Images.

B. Data Processing

CXR image preprocessing is a crucial step in supporting accurate analysis, as CXR images generally have low contrast and non-uniform light intensity distribution due to limitations of the imaging device or patient condition [34]. Gamma correction was chosen because it has previously been shown to be effective in enhancing the contrast of medical images, thus facilitating the extraction of more representative visual features [22]. Gamma correction is an image processing technique that improves image contrast by non-linearly adjusting pixel intensities [35]. Generally, the power-law method is used in the implementation of gamma correction, which can be calculated using Eq. (1) [22]:

$$I_{out} = I_{in}^{\gamma} \quad (1)$$

where I_{out} is the input pixel intensity, I_{out} is the output pixel intensity after correction, and γ (gamma) determines the level of correction applied. Fig. 2 presents the gamma correction process used to enhance the visual clarity of chest radiographs. This process begins by inputting a CXR image, after which the system calculates the average pixel brightness in the image. If the brightness value is less than 0.3, indicating that the image is too dark, the system applies a gamma value of 0.7 to increase image brightness.

Corresponding author: Ricky Eka Putra, rickyeka@unesa.ac.id, Department of Informatics, State University of Surabaya, Jl. Ketintang, Ketintang, Kec. Gayungan, Surabaya, East Java 60231, Indonesia.
Digital Object Idenfier (DOI): https://doi.org/10.35882/ijeeemi.v7i4.271

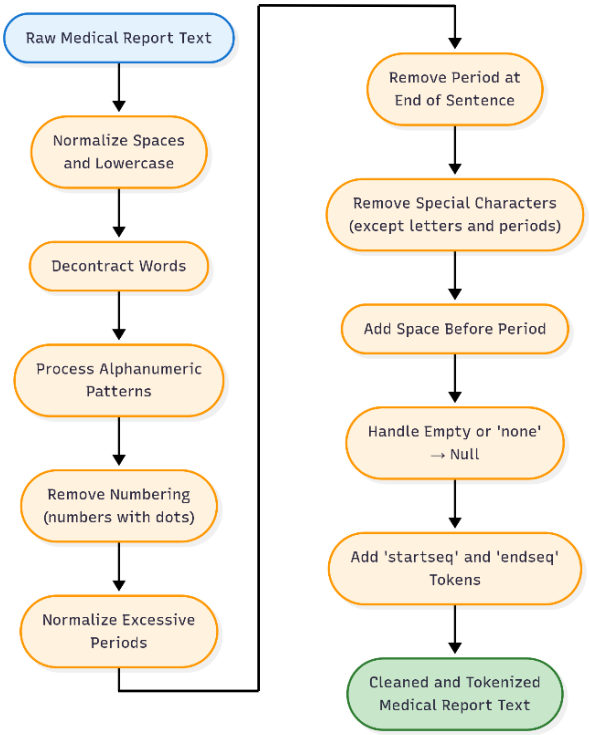


Fig. 3. Text preprocessing workflow for cleaning and standardizing medical report texts.

Conversely, if the image is sufficiently bright, the system applies a gamma value of 1.5 to adjust the contrast. This process produces a chest X-ray image that has been corrected using gamma correction. After the image quality is improved, the CXR image is resized to 224×224 pixels to meet the dimensional specifications required by CoAtNet. In addition to image preprocessing, text preprocessing is also important for noise reduction and for preparing medical reports for model training [36]. In Fig. 3, the text preprocessing workflow is performed through several stages to produce clean, standardized data ready for processing by the captioning model. These stages include normalization of spaces and letters (lowercasing), decontraction, processing of number and letter patterns, removal of numbered lists, normalization of excessive periods, and removal of periods at the end of sentences. In addition, the system removes special characters except for letters and periods, then adds a space before each period. Text entries that yield the string “none” or are empty after preprocessing are changed to None or null to indicate the absence of a description. As a final step, we incorporate the tokens “startseq” and “endseq” at the beginning and end of each text, respectively, to help the model understand sequence structure and guide text generation. To prepare textual data for model input, we employed word-level tokenization using the Keras Tokenizer, which maps each unique word in the preprocessed findings. The vocabulary size was determined from the dataset corpus, and the maximum caption length was set according to the longest sequence in the dataset. During training, word embeddings were learned end-to-end through a trainable Embedding layer with a dimensionality of 256.

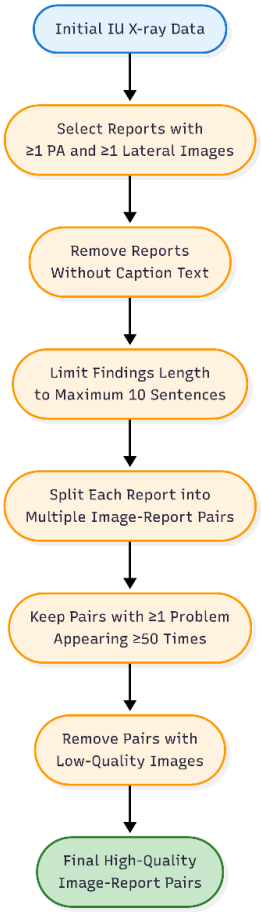


Fig. 4. Data selection workflow for generating high-quality and reliable image-caption pairs.

For image augmentation, applied exclusively to minority-label samples during the oversampling phase, we utilized offline geometric and photometric transformations, including random shifts (±5%), scaling (±5%), rotation (±15°), and brightness or contrast adjustments (±20%), implemented using the Albumentations library. Each augmented image was saved to disk before feature extraction to ensure that every synthetic image retained its original clinical caption without semantic misalignment.

Then, after text preprocessing, we performed a data selection procedure to produce high-quality image–report pairs, as shown in Fig. 4. This procedure involved selecting reports that contained at least one Posteroanterior (PA) image and one Lateral image, removing reports without caption text, limiting the findings description length to a maximum of ten sentences, and splitting each report into multiple image–report pairs according to the number of associated images. Furthermore, we retained only pairs that included at least one finding that appeared ≥50 times in the dataset, while pairs with inadequate image quality were excluded.

After the selection procedure, we partitioned the dataset into training (80%), validation (10%), and testing (10%) subsets using stratified sampling based on normal and abnormal findings, as well as PA and Lateral image projections, to ensure balanced data distribution and fair representation, as shown in Table 1. Through this stage,

Corresponding author: Ricky Eka Putra, rickyeka@unesa.ac.id, Department of Informatics, State University of Surabaya, Jl. Ketintang, Ketintang, Kec. Gayungan, Surabaya, East Java 60231, Indonesia.
Digital Object Idenfier (DOI): https://doi.org/10.35882/ijeeemi.v7i4.271

the final dataset becomes more consistent and relevant for use in model training and evaluation.

C. Handling Data Imbalance

One of the main challenges in developing medical image captioning models is the imbalanced data distribution, particularly in the IU X-Ray dataset. This imbalance can cause the model to be biased toward majority labels, making it less sensitive to rare findings. To address this issue, our study employs a methodology that reduces dominant captions while increasing underrepresented labels. This technique aims to establish a more uniform distribution of training instances, thereby facilitating balanced model learning.

It is important to note that conventional oversampling techniques such as SMOTE are unsuitable in this multimodal context. SMOTE generates synthetic samples through numerical interpolation in the feature space, producing feature vectors that lack corresponding CXR images. Since medical report generation requires strict image-text alignment, such synthetic vectors cannot be paired with original captions. Therefore, our oversampling strategy is limited to image-space augmentation, specifically geometric and photometric transformations, which preserve both the visual semantics and clinical validity of the original image–caption pairs. This ensures that augmented images remain diagnostically consistent with their ground-truth reports, allowing safe reuse of captions without introducing misalignment or semantic noise. The undersampling process reduces the dominance of overly frequent descriptions in the dataset by limiting the number of occurrences of each image projection and caption. As shown in Fig. 5, this process is applied to the training dataset after data processing by adding a new column called *projection-caption*, which combines image projection and medical finding description. The system calculates the frequency of each unique projection–caption and then applies a threshold of one occurrence for each combination. If a projection–caption appears more than once, the system randomly selects one sample; if the number is one or fewer, the system retains all samples. This process uses a random seed of 42 to ensure reproducibility and to produce a more balanced caption distribution. After the process is complete, the dominance of common captions is reduced, preventing the model from overfitting to frequently occurring descriptions.

D. Model Architecture

In this study, we define majority labels as those with an initial frequency greater than 550 and minority labels as those with a frequency lower than 250. After undersampling the majority captions, oversampling of minority labels is performed by applying a maximum frequency limit to other labels to prevent overrepresentation due to the multi-label nature of the data. As shown in Table 2, this process begins by loading the undersampled training dataset, calculating the frequency of occurrence of each label, sorting them in descending order, and then reversing the order so that the label with the lowest frequency is processed first as a priority.

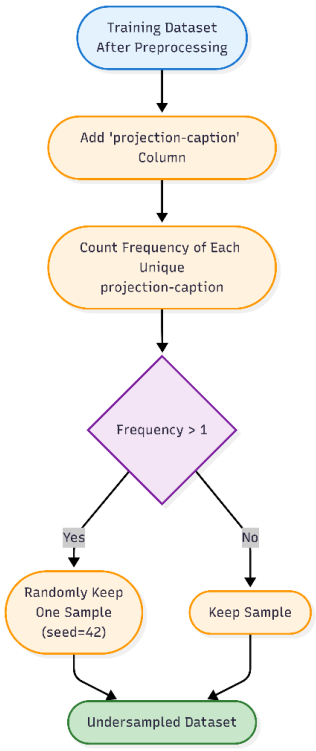


Fig. 5. Undersampling process to reduce the dominance of frequently occurring projection-caption pairs.

Two main parameters are used: max_over of 550, which serves as the maximum limit for the total frequency of a label to prevent overrepresentation, and max_num of 250, which serves as the target frequency for minority labels, since excessive augmentation of CXR images can lead to unrealistic feature learning that confuses the model [37]. Each sample is assigned an additional column, n_augs, initialized to zero to record the number of augmentations. Next, for each label, the system identifies all samples containing that label, calculates its actual frequency (including results from previous augmentations), and determines the number of additional samples needed. Only samples whose entire label set remains below the max_over threshold are selected for augmentation, and the augmentation amount is distributed proportionally among the samples to maintain a balanced data distribution. The final result of this process is a training dataset that retains the original sample size but includes an n_augs column for each image report pair, indicating the number of augmentations to be applied. Each image report pair is then augmented according to its predetermined n_augs value. As shown in Fig. 6, the architecture of the automated radiology report generation system proposed in this study is designed to generate textual descriptions from CXR images by integrating image processing, visual feature extraction, and text processing techniques. The main processes in this architecture include gamma correction, feature extraction using CoAtNet, and text processing using Bi-LSTM and MHA. The workflow begins with receiving input in the form of a CXR image, which is then processed using gamma correction to enhance contrast.

Corresponding author: Ricky Eka Putra, rickyeka@unesa.ac.id, Department of Informatics, State University of Surabaya, Jl. Ketintang, Ketintang, Kec. Gayungan, Surabaya, East Java 60231, Indonesia.
Digital Object Idenfier (DOI): https://doi.org/10.35882/ijeeemi.v7i4.271

Table 2. Pseudocode for multi-label oversampling algorithm to balance minority label frequencies.

Oversampling	
Input:	
train_data	Dataset after undersampling based on majority captions
max_over	Maximum allowed total frequency for any label (e.g., 550)
max_num	Target minimum frequency for minority labels (e.g., 250)
Output:	Training dataset with n_augs field indicating augmentation count per sample
	START
Step 1: Compute Frequency	tags, tag_counter, tag_df = ComputeFrequency(train_data) train_tags = SortLabelsDescending(tag_df['Tag']) train_tags.reverse()
Step 2: Initialize Augmentation Counter	// Add a new key 'n_augs' to track how many times this image-report pair will be augmented For each sample d in train_data: d['n_augs'] = 0
Step 3: Iterate Over Each projection-caption	For each (caption, count) in frequencies: filtered_data = [d for d in new_train if d['projection-caption'] == caption] If count ≤ max_samples: sampled_items = filtered_data Else: generator = RandomGenerator(seed) sampled_items = generator.sample(filtered_data, max_samples) undersampled_train = undersampled_train + sampled_items

Step 4: Iterate Over Each Label (from rarest to most common)	For each tag in train_tags:
	// Find all samples containing this probelm label (tag)
	partial_ids = [index of d in train_data where tag in tags of d]
	// Compute current total frequency of this label
	current_tag_count = Sum(train_data[idx]['n_augs'] for idx in partial_ids) + Length(partial_ids)
	num_sampling = max_num - current_tag_count
	// If current frequency < max_num, we need to augment; otherwise, skip
	available_partial_ids = []
	For each idx in partial_ids:
	If not IsOver(train_data[idx][tag_key]):
	Append idx to available_partial_ids
	// Only allow augmentation if no label in the sample exceeds max_over
	// Distribute augmentation proportionally among available samples
	If num_sampling > 0:
	base_value = IntegerDivision(num_sampling, Length(available_partial_ids))
	remainder = num_sampling % Length(available_partial_ids)
	For i = 0 to Length(available_partial_ids)-1:
	If i < remainder:
	train_data[available_partial_ids[i]]['n_augs'] += base_value + 1
	Else:
	train_data[available_partial_ids[i]]['n_augs'] += base_value

	END

Corresponding author: Ricky Eka Putra, rickyeka@unesa.ac.id, Department of Informatics, State University of Surabaya, Jl. Ketintang, Ketintang, Kec. Gayungan, Surabaya, East Java 60231, Indonesia.
Digital Object Idenfier (DOI): https://doi.org/10.35882/ijeeemi.v7i4.271

This method is essential because it improves pixel intensity distribution, making the image clearer and more informative for the subsequent feature extraction process.

We formalize the task as supervised sequence-to-sequence learning: given an input image $x \in \mathbb{R}^{H\times W\times C}$, the model generates a radiology report $y_{1:T} = (y_1, \ldots, y_T)$, where each y_t is a token from a fixed medical vocabulary. The joint distribution is factorized autoregressively as Eq. (2) [38]:

$$p(y_{1:T} \mid x) = \prod_{t=1}^{T} p(y_t \mid y_{<t}, Z) \quad (2)$$

where $Z = \{z_1, \ldots, z_L\}$is a sequence of visual features extracted from $x$by the encoder, and $y_{<t}$denotes previously generated tokens. The decoder estimates each conditional distribution using a softmax over the vocabulary like in Eq. (3) [38]:

$$p(y_t \mid y_{<t}, Z) = \mathrm{Softmax}(W_o s_t + b_o)_{y_t} \quad (3)$$

with $s_t$being the fused decoder state at step t.

After going through the gamma correction stage, the X-ray image is provided as input to CoAtNet. CoAtNet is a hybrid architecture that combines convolutional layers and self-attention mechanisms to improve the quality of feature extraction across various data scenarios [29]. This architecture was designed to address the limitations of CNN and transformer models, particularly in handling large-scale data. In CoAtNet's design, convolutional layers excel at capturing local spatial features, whereas the self-attention mechanism excels at establishing global connections among pixels. The combination of these two components enables the model to effectively capture complex image structures without losing important information, resulting in richer visual representations for the medical report generation process. Formally, CoAtNet processes $x \in \mathbb{R}^{H\times W\times C}$ through stacked stages of convolution ($\mathcal{C}$) and relative-position self-attention ($\mathcal{A}$). In the early stages, depthwise separable convolutions are applied to introduce local inductive bias and perform downsampling, as shown in Eq. (4) [29]:

$$X^{(l+1)} = \mathcal{C}^{(l)}(X^{(l)}) \quad (4)$$

Later stages use self-attention with feed-forward networks Eq. (5):

$$X^{(l+1)} = FFN(\mathcal{A}^{(l)}(X^{(l)})) + X^{(l)} \quad (5)$$

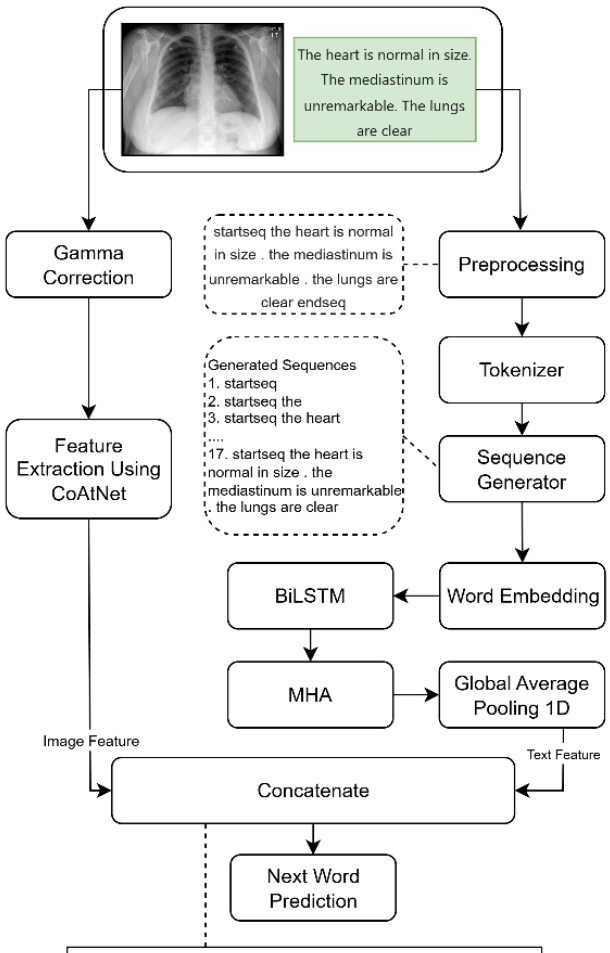
where attention is computed as Eq. (6) [29]:

$$\mathcal{A}(X) = softmax(\frac{QK^{\top}}{\sqrt{d_k}} + B)V \quad (6)$$

with B denoting a learnable relative positional bias. The final activation map $X^{(L)}$ is globally average-pooled to produce a compact visual representation, Eq. (7) [29]:

$$Z = Pool(X^{(L)}) \in \mathbb{R}^{L\times d} \quad (7)$$

CoAtNet was selected over other state-of-the-art encoders due to its demonstrated superiority in multi-label chest X-ray classification tasks. In a recent comprehensive benchmark conducted on the NIH ChestX-ray14 dataset [30], CoAtNet achieved the highest area under the receiver operating characteristic (AUROC) score of 84.2% among all individual models evaluated, including ConvNeXtV2 (84.1%), Swin Transformer V2 (83.6%), and DenseNet-121 (82.4%). This superior classification performance indicates that CoAtNet generates more discriminative and semantically relevant visual feature vectors.



startseq the heart is normal in size . the mediastinum is unremarkable . the lungs are clear endseq

START

Features	Target
[Image Features] + [Text Features of "startseq"]	the
[Image Features] + [Text Features of "startseq the"]	heart
[Image Features] + [Text Features of "startseq the heart"]	is
…	…
[Image Features] + [Text Features of "startseq the heart is normal in size . the mediastinum is unremarkable . the lungs are"]	clear
[Image Features] + [Text Features of "startseq the heart is normal in size . the mediastinum is unremarkable . the lungs are clear"]	endseq

END

Fig. 6. Proposed architecture for automatic CXR report generation using CoAtNet and Bi-LSTM with MHA.

Such high-quality feature representations are essential for downstream tasks, including automated radiology report generation. Fig. 7 presents the CoAtNet architecture used as a feature extractor in this study. We employ the CoAtNet-2 variant, which consists of two convolutional stem layers (S0), followed by two MBConv blocks in S1 (56×56), six in S2 (28×28), fourteen relative attention blocks in S3 (14×14), and two in S4 (7×7), with channel dimensions of 128, 128, 256, 512, and 1024, respectively.

Corresponding author: Ricky Eka Putra, rickyeka@unesa.ac.id, Department of Informatics, State University of Surabaya, Jl. Ketintang, Ketintang, Kec. Gayungan, Surabaya, East Java 60231, Indonesia.
Digital Object Idenfier (DOI): https://doi.org/10.35882/ijeeemi.v7i4.271

This model receives a 224×224-pixel CXR image as input and processes it through a stem stage comprising a 3×3 convolution with a stride of 2, followed by another 3×3 convolution for downsampling.

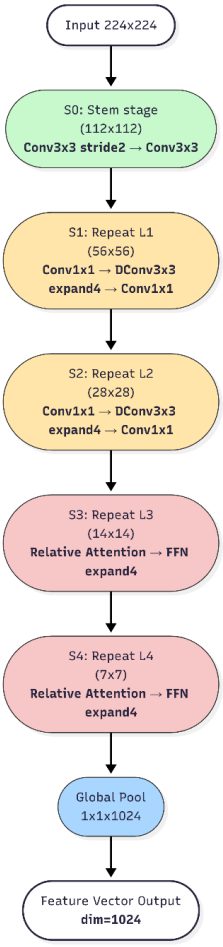


Fig. 7. Pretrained CoAtNet model serves as a feature extractor for CXR images.

After the initial stage, the image passes through the MBConv block at stages S1 (56×56) and S2 (28×28), which captures local patterns through a combination of pointwise convolution, depthwise convolution with a fourfold expansion, and pointwise projection. The subsequent stages, S3 (14×14) and S4 (7×7), utilize relative attention combined with a fourfold expansion feed-forward network to represent global relationships among features. After global average pooling, a 1024-dimensional vector representation is produced, which serves as the visual feature input for the captioning process.

In this study, CoAtNet was pretrained on the NIH ChestX-ray14 dataset [30], a large-scale medical imaging benchmark containing 112,120 CXRs labeled for 14 thoracic diseases. This domain-specific pretraining ensures that the extracted features are clinically relevant and can be effectively transferred to the report generation task using the smaller IU X-ray dataset.

In parallel, the model also processes the radiology report text. The reference text (ground truth) first undergoes preprocessing to remove noise and standardize the format. Afterward, the text is tokenized into a sequence of tokens, which are then converted into word embeddings using a text embedding model. The vector representation of each word is subsequently processed by a Bi-LSTM to capture the contextual relationships across the entire word sequence. Bi-LSTM was chosen for its ability to process information in both directions forward and backward allowing it to capture the context of preceding and succeeding words in a sentence. This capability enhances semantic comprehension and improves the quality of text feature extraction [39]. The decoder maintains forward and backward LSTM states over the generated (or teacher-forced) embeddings e_t. The forward pass is computed as shown in Eq. (8) [40]:

$$\begin{aligned} i_t &= \sigma(W_i e_t + U_i h_{t-1} + b_i), \\ f_t &= \sigma(W_f e_t + U_f h_{t-1} + b_f), \\ o_t &= \sigma(W_o' e_t + U_o h_{t-1} + b_o'), \\ \tilde{c}_t &= \tanh(W_c e_t + U_c h_{t-1} + b_c), \\ c_t &= f_t \odot c_{t-1} + i_t \odot \tilde{c}_t, \\ \vec{h}_t &= o_t \odot \tanh(c_t). \end{aligned} \quad (8)$$

A symmetric backward LSTM yields $\overleftarrow{h}_t$. The decoder state is the concatenation $h_t = [\vec{h}_t; \overleftarrow{h}_t]$. Before prediction, $h_t$queries the encoder via MHA to produce a context vector a_t. The fused state $s_t = \phi([h_t; a_t])$ (with ϕ a feed-forward layer and normalization) conditions the token distribution via a projection and softmax.

The next step involves applying MHA to enhance the model's capacity to capture complex contextual representations in sequential data [39]. MHA is an extension of self-attention or more precisely, scaled dot-product attention which calculates attention weights based on the interactions among three main components: *Q* (Query), *K* (Key), and *V* (Value). The basic formula of this mechanism is shown in Eq. (9) [22]

$$Attention(Q,K,V) = softmax\left(\frac{Q\,K^T}{\sqrt{d_k}}\right)V \quad (9)$$

where d_k indicates the dimensionality of the Key matrix, which plays a key role in calculating the attention score. This mechanism allows each token in the input sequence to distribute its focus or attention weights to other tokens within the same sequence. To expand its representational capacity, this mechanism is extended to a multi-head structure, allowing the attention process to operate across multiple attention heads in parallel. Each head has unique projection parameters that generate distinct sets of *Q*, *K*, and *V* to capture diverse relational patterns in sequential data.

After each head generates a contextual representation, the results are combined through a concatenation operation and projected back into the final representation space using the transformation matrix W^O. The mathematical formulation of MHA is presented in Eq. (10) and Eq. (11) [22].

$$MultiHead(Q,K,V) = [head_1, \ldots, head_h]W^O \quad (10)$$

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \quad (11)$$

Corresponding author: Ricky Eka Putra, rickyeka@unesa.ac.id, Department of Informatics, State University of Surabaya, Jl. Ketintang, Ketintang, Kec. Gayungan, Surabaya, East Java 60231, Indonesia.
Digital Object Idenfier (DOI): https://doi.org/10.35882/ijeeemi.v7i4.271

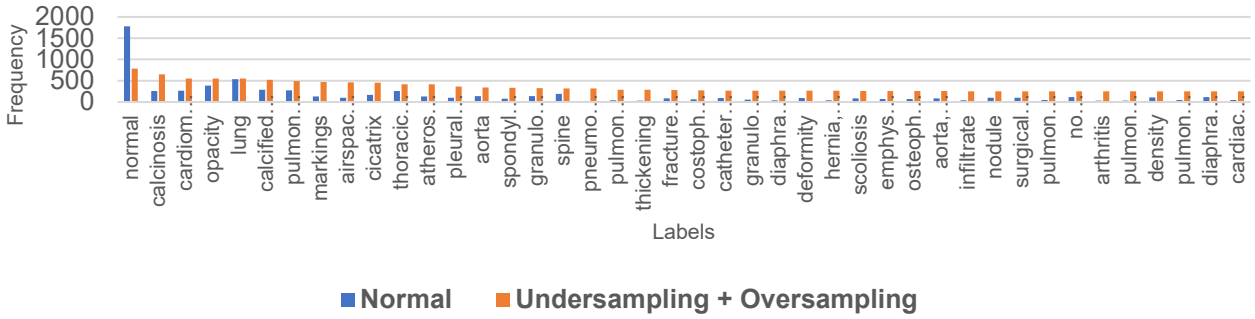


Fig. 8. Comparison of label frequency before and after applying undersampling and oversampling.

Table 3. Comparative performance evaluation of our proposed models under different sampling strategy.

Sampling	Model	B-1	B-2	B-3	B-4	R-L	Total
No Sampling	CoAtNet + LSTM-MHA	**0.481**	0.309	0.218	0.155	0.368	1.530
Undersampling	CoAtNet + LSTM-MHA	0.448	0.311	0.233	0.175	0.411	1.579
Oversampling	CoAtNet + LSTM-MHA	0.444	0.318	0.238	0.175	0.413	1.588
Undersampling + Oversampling	**CoAtNet + LSTM-MHA**	0.474	**0.324**	**0.244**	**0.182**	**0.394**	**1.618**

Table 4. Performance comparison of different encoder architectures using the same data and decoder.

Encoder	B-1	B-2	B-3	B-4	R-L	Total
CheXNet [22]	0.453	0.302	0.205	0.143	0.354	1.457
Swin Transformer V2 [44]	0.422	0.290	0.214	0.158	0.386	1.470
ConvNeXt V2 [45]	0.456	0.295	0.208	0.146	0.368	1.474
ViT [46]	0.425	0.286	0.213	0.160	**0.400**	1.485
CoAtNet [29]	**0.474**	**0.324**	**0.244**	**0.182**	0.394	**1.618**

In our implementation, the Multi-Head Attention mechanism is applied after the Bi-LSTM layer, using the Bi-LSTM output as the input for the Q, K, and V projections. Specifically, the MHA module consists of eight parallel attention heads, each operating on a projected subspace with a dimensionality equal to the word embedding size (256). This configuration enables the model to jointly attend to information from different representation subspaces at various positions in the sequence. Following the MHA computation, a residual connection is added between the Bi-LSTM output and the MHA output, followed by layer normalization to stabilize training dynamics and preserve gradient flow.

After attention, the resulting sequence is globally pooled using Global Average Pooling to produce a fixed-length textual representation, which is then concatenated with the visual feature vector from CoAtNet. This integration allows the decoder to condition word prediction on both the bidirectional contextual semantics from Bi-LSTM with MHA and the rich visual features from CoAtNet.

This process ensures that contextual information from multiple heads can be optimally integrated into the final representation. Subsequently, the CoAtNet-derived visual features and Bi-LSTM-MHA processed textual features are merged via a concatenation layer. This combined feature vector is then passed through a dense layer responsible for predicting the next word in the radiology report. The model generates words sequentially, producing one token per iteration until the complete report is constructed.

E. Training and Evaluation Details

The entire training process was executed using an NVIDIA GeForce RTX 4060 graphics processing unit (8 GB video RAM) on Windows 11 with WSL2 (Ubuntu 22.04). We configured 15 epochs without early stopping and a batch size of 2, choices determined by GPU memory constraints. These small epoch and batch size values are consistent with research practices [18], [19], [22], [26] to address the high complexity and large memory requirements of similar models.

This model employs the AdamW optimizer with an initial learning rate of 0.001, which is dynamically reduced during training using the ReduceLROnPlateau scheduler (factor = 0.5, patience = 4, minimum learning rate = 1×10^{-8}). The LeakyReLU activation function (with a

Corresponding author: Ricky Eka Putra, rickyeka@unesa.ac.id, Department of Informatics, State University of Surabaya, Jl. Ketintang, Ketintang, Kec. Gayungan, Surabaya, East Java 60231, Indonesia.
Digital Object Idenfier (DOI): https://doi.org/10.35882/ijeeemi.v7i4.271

(a) (b) (c)

Fig. 9. High-similarity of generated medical reports versus ground truth reference captions. (a) CXR3302_IM-1579-1001, (b) CXR3925_IM-1999-1003002, (c) CXR58_IM-2177-2001.

(a) CXR3302_IM-1579-1001:
- Ground Truth: “the heart is normal in size . the mediastinum is unremarkable . the lungs are slightly hypoinflated but clear . there is no pleural effusion . no acute disease”
- Prediction: “the heart is normal in size. the mediastinum is unremarkable. the lungs are clear. no focal consolidation pneumothorax or pleural effusion. no acute cardiopulmonary abnormality”
- Metrics: B-1 = 0.7931, B-2 = 0.7336, B-3 = 0.6858, B-4 = 0.6437, R-L = 0.7931, Total = 3.6494.

(b) CXR3925_IM-1999-1003002:
- Ground Truth: “the lungs are clear . there is no pleural effusion or pneumothorax . the heart is not significantly enlarged . there are calcified right hilar and mediastinal lymph there are atherosclerotic changes of the aorta . arthritic changes of the skeletal structures are noted . no acute pulmonary disease"
- Prediction: “the heart is normal in size . the mediastinum is unremarkable . the lungs are clear . there is no pleural effusion or pneumothorax . the heart is not significantly enlarged . there are atherosclerotic changes of the spine . emphysematous changes of the skeletal structures are noted . no acute pulmonary disease”
- Metrics: B-1 = 0.7358, B-2 = 0.7137, B-3 = 0.7002, B-4 = 0.6774, R-L = 0.7647, Total = 3.5919.

(c) CXR58_IM-2177-2001:
- Ground Truth: “the heart is normal in size . the mediastinum is unremarkable . the lungs are clear . mild scoliosis and degenerative changes of the thoracic spine noted . no acute disease”
- Prediction: “the heart is normal in size . the mediastinum is unremarkable . the lungs are clear . no focal consolidation pneumothorax or pleural effusion . emphysematous changes are noted . no acute cardiopulmonary abnormality”
- Metrics: B-1 = 0.6471, B-2 = 0.6104, B-3 = 0.5859, B-4 = 0.5563, R-L = 0.6769, Total = 3.0766.

negative slope coefficient of 0.3) is used to mitigate the dying ReLU problem [41]. The Categorical Crossentropy loss function is selected because each word prediction step is treated as a multiclass classification task over the target vocabulary, a standard approach in sequence-to-sequence report generation [18]. To mitigate overfitting, a Dropout layer with rate 0.5 is applied after the final dense layer, which randomly deactivates half of the neuronal units during model training [41].

In this study, the present investigation employed BLEU and ROUGE-L as quantitative measures for evaluating caption quality. BLEU functions as a precision-oriented metric that quantifies n-gram correspondence between system-generated descriptions and reference texts, thus evaluating fluency and grammatical accuracy [42]. Meanwhile, ROUGE-L measures similarity based on Longest Common Subsequence (LCS), which allows for assessment of information completeness and semantic coherence despite variations in sentence structure [43]. The combination of these two metrics provides a balanced and relevant evaluation for medical image captioning tasks. At the end of each epoch, the system saves the model and evaluates it against all metrics, namely BLEU-1 (B-1), BLEU-2 (B-2), BLEU-3 (B-3), BLEU-4 (B-4), and ROUGE-L (R-L), to identify the optimal model based on the highest cumulative score.

III. Results

A. Label Distribution After Sampling

Before the sampling technique was implemented, the training dataset suffered from extreme imbalance, with majority labels such as normal dominating (1776 occurrences), while minority labels such as pneumonia appeared only 28 times, risking model bias toward the majority class and low sensitivity to abnormal findings. To address this, a combined method of undersampling the majority captions and oversampling the minority labels was utilized, successfully balancing the label distribution, as shown in Fig. 8, by reducing the frequency of normal to 788 and significantly increasing the frequency of minority labels, such as pneumonia from 28 to 319 and pulmonary edema from 38 to 289, as well as dozens of other labels such as cardiomegaly and opacity. This created a fairer and more representative training

Corresponding author: Ricky Eka Putra, rickyeka@unesa.ac.id, Department of Informatics, State University of Surabaya, Jl. Ketintang, Ketintang, Kec. Gayungan, Surabaya, East Java 60231, Indonesia.
Digital Object Idenfier (DOI): https://doi.org/10.35882/ijeeemi.v7i4.271

(a) (b) (c)

Fig. 10. Low-similarity of generated medical reports compared to ground truth reference captions. (a) CXR3390_IM-1636-2001, (b) CXR2969_IM-1360-1001, (c) CXR3111_IM-1461-2001.

(a) CXR3390_IM-1636-2001:
- Ground Truth: “the lungs are clear . there is no focal airspace consolidation . no pleural effusion or pneumothorax . normal cardiomediastinal silhouette . no evidence of active disease”
- Prediction: “the heart is normal in size . the mediastinum is unremarkable . the lungs are clear . there is no focal consolidation pleural effusion or pneumothorax . the heart is not significantly enlarged . there are atherosclerotic changes of the spine . emphysematous changes of the skeletal structures are noted . no acute pulmonary disease”
- Metrics: B-1 = 0.3818, B-2 = 0.3032, B-3 = 0.2624, B-4 = 0.2273, R-L = 0.4390, Total = 1.6137.

(b) CXR2969_IM-1360-1001:
- Ground Truth: “heart size and mediastinal contours are stable . atherosclerotic calcifications of the aorta . moderate severe hyperexpansion of the lungs and decreased peripheral vascular markings consistent with emphysema . stable biapical pleural parenchymal scarring . scattered granulomas . no abnormal airspace consolidation . no pneumothorax or pleural effusion . stable emphysematous changes . stable biapical pleural parenchymal scarring”
- Prediction: “the heart is normal in size . the mediastinum is unremarkable . the lungs are otherwise clear . there is no focal consolidation pleural effusion or pneumothorax . the heart is not significantly enlarged . there are atherosclerotic changes of the aorta . emphysematous changes of the skeletal structures are noted . no acute pulmonary disease”
- Metrics: B-1 = 0.4480, B-2 = 0.2508, B-3 = 0.1336, B-4 = 0.0800, R-L = 0.2807, Total = 1.1930.

(c) CXR3111_IM-1461-2001:
- Ground Truth: “stable appearance of previous sternotomy . stable cardiomegaly . stable mild bilateral interstitial opacities in which may represent mild pulmonary edema . no evidence of large pleural effusion or pneumothorax . stable cardiomegaly and mild bilateral interstitial opacities which represent mild pulmonary edema”
- Prediction: “the heart is normal in size . the mediastinum is unremarkable . the lungs are clear without focal consolidation pleural effusion or pneumothorax . there is mild degenerative changes of the thoracic spine . no acute cardiopulmonary abnormality”
- Metrics: B-1 = 0.2769, B-2 = 0.1811, B-3 = 0.1364, B-4 = 0.1047, R-L = 0.1975, Total = 0.8965.

environment, allowing the model to learn from a variety of clinical findings in a balanced manner.

B. Sampling Strategy on Model Performance

The implementation of the sampling strategy significantly improved the scores of the CoAtNet + LSTM-MHA model in generating automated medical reports, as seen in Table 3, where the baseline model without sampling achieved only a cumulative score of 1.530. Undersampling increased the R-L score from 0.368 to 0.411, indicating improved completeness and semantic structure. Although B-1 decreased slightly, the increases in B-2, B-3, and B-4 evidenced more accurate and varied descriptions. Oversampling slightly outperformed with a cumulative score of 1.588, driven by consistent improvements in B-2, B-3, and R-L. However, the combination of undersampling and oversampling proved most effective, achieving the highest cumulative score of 1.618, as it was able to reduce majority class bias while enriching minority class representation, thus proving that addressing data imbalance through sampling strategies is a decisive factor in improving the quality of model-generated medical reports.

C. Comparison of Encoder Architectures

To evaluate the superiority of CoAtNet as a visual encoder, its performance was benchmarked against four state-of-the-art encoders using a dataset balanced via undersampling and oversampling techniques alongside an identical LSTM-MHA decoder, ensuring that the performance difference purely reflects visual feature extraction capability. As shown in Table 4, CoAtNet

Corresponding author: Ricky Eka Putra, rickyeka@unesa.ac.id, Department of Informatics, State University of Surabaya, Jl. Ketintang, Ketintang, Kec. Gayungan, Surabaya, East Java 60231, Indonesia.
Digital Object Idenfier (DOI): https://doi.org/10.35882/ijeeemi.v7i4.271

recorded the highest cumulative score (1.618), significantly outperforming ViT (1.485), ConvNeXt V2 (1.474), Swin Transformer V2 (1.470), and CheXNet (1.457). CoAtNet also demonstrated consistent improvement across almost all metrics, with higher B-1 to B-4 scores, indicating its superior ability to generate structurally and linguistically precise medical descriptions, although ViT showed a slight advantage in R-L. These findings confirm that CoAtNet's hybrid methodology, integrating CNN's local feature extraction capabilities with Transformer's attention mechanism, delivers superior performance in producing CXR-based medical reports compared with conventional pure CNN or Transformer frameworks.

Fig. 11. Generated reports using different sampling strategies compared to ground truth (CXR2368_IM-0928-2001).

CXR2368_IM-0928-2001:

- Ground Truth: “the heart and mediastinum are unremarkable . the lungs are clear without infiltrate . there is no effusion or pneumothorax . there is mild degenerative changes of the thoracic spine . no acute cardiopulmonary disease”
- Undersampling: “the heart is normal in size . the mediastinum is unremarkable . the lungs are clear . no focal consolidation . no pneumothorax or pleural effusion . no acute cardiopulmonary abnormality
- Oversampling: the heart size and mediastinal contours are within normal limits . the lungs are clear . no focal consolidation pneumothorax or pleural effusion . no acute bony abnormality . no acute cardiopulmonary abnormality”
- Combined: “the heart is normal in size . the mediastinum is stable . the lungs are clear without evidence of acute infiltrate or effusion . there is no evidence of pneumothorax . there are mild degenerative changes of the thoracic spine . no acute cardiopulmonary disease”

D. Model Performance with Bi-LSTM

A consistent improvement in model performance was observed when the decoder architecture was upgraded from LSTM to Bi-LSTM, as shown in Table 5. Using the same dataset and encoder, replacing the one-way LSTM with a bidirectional Bi-LSTM successfully increased the cumulative score from 1.618 to 1.642. This improvement can be seen from the increase in values across all evaluation metrics: B-1 rose from 0.474 to 0.480, B-2 from 0.324 to 0.329, B-3 from 0.244 to 0.245, B-4 from 0.182 to 0.183, and R-L from 0.394 to 0.405. These findings demonstrate that the Bi-LSTM’s ability to capture semantic context from both directions (forward and backward) strengthens the model’s understanding of word relationships, resulting in medical reports that more closely resemble the original references.

Table 5. Comparative performance metrics between LSTM and Bidirectional (Bi) LSTM on Medical Report Generation.

Bi	B-1	B-2	B-3	B-4	R-L	Total
✗	0.474	0.324	0.244	0.182	0.394	1.618
✓	**0.480**	**0.329**	**0.245**	**0.183**	**0.405**	**1.642**

E. Generated Reports

Fig. 9 shows that the model demonstrated a high capability in generating medical reports that were close in structure and meaning to the original reports, especially in simple cases. In several image examples, the model successfully captured key clinical findings such as a normal-sized heart, an unaffected mediastinum, and the absence of pleural effusion, although it sometimes simplified or slightly modified the descriptions, such as substituting phrases related to lung conditions or replacing the term degenerative spine with other relevant terms. Despite these minor deviations, the cumulative score remained high (up to 3.6494), indicating that the model consistently retained the core clinical information. These results demonstrate the model’s reliability in reproducing key findings in non-complex radiology cases.

Fig. 10 shows that the model tends to simplify complex clinical findings and fails to capture important details from the original report, sometimes even adding irrelevant information. In some cases, the model misses key descriptions such as normal cardiomediastinal silhouette, pulmonary hyperinflation, or stable cardiomegaly, and omits important qualifiers such as “stable,” resulting in a cumulative score drop to 0.8965. These results indicate that the model still struggles to handle cases with multiple abnormalities or specific clinical terminology that rarely appears in the training data.

IV. Discussion

It is important to note that our task is medical image captioning, not multi-label classification. The labels in the IU X-ray dataset are used only to assist in data balancing for image-text pairs with minority labels, not as direct prediction targets. Therefore, model performance cannot be evaluated using per-class classification metrics such as precision, recall, or F1-score.

A. The Impact of Sampling and Model Architecture

The combined undersampling and oversampling strategy significantly improved model performance in generating fairer and more accurate medical reports by balancing the label frequency distribution, where the dominance of majority labels such as normal was reduced from 1.776 to

Corresponding author: Ricky Eka Putra, rickyeka@unesa.ac.id, Department of Informatics, State University of Surabaya, Jl. Ketintang, Ketintang, Kec. Gayungan, Surabaya, East Java 60231, Indonesia.
Digital Object Idenfier (DOI): https://doi.org/10.35882/ijeeemi.v7i4.271

788, while the representation of minority labels such as pneumonia was increased from 28 to 319, as seen in Fig. 8. This distribution transformation created a more representative training environment, which was reflected in the increase in the model's cumulative score from 1.530 without sampling to 1.618 with the combination of techniques, as shown in Table 3, where the improvement occurred in both linguistic precision (BLEU) and semantic completeness (ROUGE-L) metrics. In addition to the data balancing strategy, the superiority of CoAtNet as a visual encoder also strengthened model performance by extracting richer and more contextual visual features from the balanced CXR images, as it combines convolutional layers for spatial detail and a self-attention mechanism for global contextualization, overcoming the limitations of conventional CNNs and pure Transformers. The relevance of these features is reinforced by pretraining on the NIH ChestX-ray14 dataset, which makes the model more sensitive to pathological chest patterns. As shown in Table 4, CoAtNet consistently outperforms four other state-of-the-art encoders with the highest cumulative score of 1.618, demonstrating the synergy between data balancing and hybrid feature extraction as the foundation of the system's high performance.

Fig. 13. Comparative report generation outputs for a CXR image using LSTM and Bi-LSTM (CXR3925_IM-1999-1003002).

CXR3925_IM-1999-1003002:

- Ground Truth: "the lungs are clear . there is no pleural effusion or pneumothorax . the heart is not significantly enlarged . there are calcified right hilar and mediastinal lymph . there are atherosclerotic changes of the aorta . arthritic changes of the skeletal structures are noted . no acute pulmonary disease"
- LSTM: "the heart is normal in size . the mediastinum is unremarkable . the lungs are clear . there is no pleural effusion or pneumothorax . there is no pleural effusion . no acute cardiopulmonary disease"
- Bi-LSTM: "the heart is normal in size . the mediastinum is unremarkable . the lungs are clear . there is no pleural effusion or pneumothorax . the heart is not significantly enlarged . there are atherosclerotic changes of the spine . emphysematous changes of the skeletal structures are noted . no acute pulmonary disease"

(a) (b)

Fig. 12. Generated reports of minority labels such as pneumonia and pulmonary emphysema. (a) Pneumonia (CXR1157_IM-0106-1001), (b) Pulmonary emphysema (CXR3354_IM-1609-1001).

(a) CXR1157_IM-0106-1001:

- Ground Truth: "the heart pulmonary and mediastinum are within normal limits . there is no pleural effusion or pneumothorax . there is a region of left upper lobe perihilar opacity identified . left upper lobe pneumonia . followup radiographs after appropriate therapy in weeks are indicated to exclude an underlying abnormality"
- Prediction: "the heart is normal in size . the mediastinum is unremarkable . the lungs are clear . there is no focal consolidation pleural effusion or pneumothorax . there is an acute cardiopulmonary disease"

(b) CXR3354_IM-1609-1001:

- Ground Truth: "heart size and mediastinal contours appear within normal limits . hyperinflated lungs with flattening of diaphragms compatible with emphysema . no focal consolidation pleural effusion or pneumothorax . no acute bony abnormality . emphysema . no acute cardiopulmonary abnormality"
- Prediction: "the heart is normal in size . the mediastinum is unremarkable . the lungs are clear . no focal consolidation pneumothorax or pleural effusion . emphysematous changes are noted . no acute cardiopulmonary abnormality"

To further illustrate the nuanced impact of each sampling strategy, qualitative analysis across clinical scenarios, as exemplified in Fig. 11, reveals that undersampling alone enhances structural coherence and reduces overuse of generic phrasing, yet may omit subtle pathological descriptors due to insufficient exposure to rare findings. In contrast, oversampling improves the model's sensitivity to minority conditions, such as degenerative or chronic changes, by enriching their representation, though it occasionally introduces minor

Corresponding author: Ricky Eka Putra, rickyeka@unesa.ac.id, Department of Informatics, State University of Surabaya, Jl. Ketintang, Ketintang, Kec. Gayungan, Surabaya, East Java 60231, Indonesia.
Digital Object Idenfier (DOI): https://doi.org/10.35882/ijeeemi.v7i4.271

Table 6. Comparative performance evaluation of our proposed model with previous research methods.

Model	B-1	B-2	B-3	B-4	R-L	Total
ResNet-152 & LSTM with Multi-level attention [19]	-	-	-	0.125	0.262	-
ResNet-50 + CVAM and LSTM [20]	0.460	0.294	0.207	0.152	0.385	1.498
Swin Transformer and BERT [25]	-	-	-	0.151	0.343	-
CNX-B2: ConvNeXt and BioBERT [26]	0.479	**0.363**	**0.261**	0.173	0.354	1.630
Ours	**0.480**	0.329	0.245	**0.183**	**0.405**	**1.642**

redundancy or overgeneralization. Only the combined approach consistently preserves both the precision of common findings and the fidelity of infrequent abnormalities. Furthermore, performance improvements were reinforced by the Bi-LSTM architecture, which enhances the model's capacity through bidirectional (forward–backward) word sequence processing, in contrast to unidirectional LSTMs that risk losing important information at the end of sentences. This capability allows for a more accurate capture of semantic relationships between phrases. As shown in Table 5, replacing the LSTM with the Bi-LSTM increased the cumulative score from 1.618 to 1.642, with consistent improvements across all metrics, including R-L, which rose from 0.394 to 0.405, indicating improved completeness and semantic coherence of the reports. This integration of CoAtNet, sampling, and encoder strategies forms a pipeline that yields a superior medical report generation system.

The ability of Bi-LSTM to understand context from both directions enhances the accuracy of clinical descriptions. In Fig. 12, the LSTM-based model tends to overlook critical details toward the end of the report, such as atherosclerotic changes and arthritic bone structures, and excessively repeats phrases such as "no pleural effusion." In contrast, the Bi-LSTM successfully captures and reorganizes this information more comprehensively and coherently, despite minor terminological inaccuracies. This demonstrates that bidirectional processing enables the model to consider phrase-to-phrase relationships in a more holistic manner, resulting in reports that better reflect the structure and meaning of the original reference.

To further validate the model's clinical sensitivity to minority pathologies, we conducted a qualitative examination of generated reports for cases with low-frequency abnormalities, such as pneumonia (frequency: 28) and pulmonary emphysema (frequency: 32). As shown in Fig. 13, the model successfully captured and articulated key diagnostic phrases. In the emphysema case, the model correctly generated "emphysematous changes are noted" while preserving the full clinical context, including normal heart size, clear lungs, and absence of acute abnormalities. In the pneumonia case, however, the model failed to identify the key pathological finding (left upper lobe opacity) and incorrectly reported "the lungs are clear," which contradicts the ground truth. Although it correctly noted the absence of pleural effusion or pneumothorax, the omission of pneumonia and the erroneous assertion of clear lungs represent a clinically significant error.

B. Comparison with Previous Works

Our proposed model achieved the highest quantitative performance with a cumulative score of 1.642, significantly outperforming previous state-of-the-art approaches such as CNX-B2 (1.630), ResNet-50+CVAM (1.498), and the Swin Transformer+BERT and ResNet-152+LSTM architectures. This superiority is reflected in the improvement of the R-L metric to 0.405 and the superior balance between linguistic precision (BLEU) and semantic completeness (ROUGE-L), although CNX-B2 excels in B-2 and B-3. As shown in Table 6, our approach not only produces more linguistically precise descriptions but also clinically fairer ones through better representation of minority findings. A key contribution lies in our comprehensive solution, which demonstrates that performance improvements stem from addressing the root cause of data imbalance rather than merely increasing model complexity.

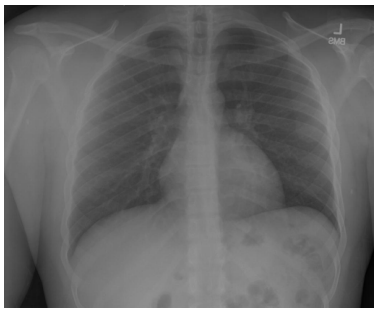


Fig. 14. Qualitative comparison of generated report between ours and previous method (CXR52_IM-2131-1001).

CXR52_IM-2131-1001:

- Ground Truth: "the heart is normal size . the mediastinum is unremarkable . there is no pleural effusion pneumothorax or focal airspace disease . the are unremarkable . no acute cardiopulmonary abnormality"
- CNX-B2 [26]: "no acute cardiopulmonary abnormality. the lungs are clear bilaterally. specifically, no evidence of focal consolidation, pneumothorax, or pleural effusion. cardiomediastinal silhouette is un"
- Ours: "the heart is normal in size . the mediastinum is unremarkable . the lungs are clear . no focal consolidation pneumothorax or pleural effusion . no acute cardiopulmonary abnormality"

Corresponding author: Ricky Eka Putra, rickyeka@unesa.ac.id, Department of Informatics, State University of Surabaya, Jl. Ketintang, Ketintang, Kec. Gayungan, Surabaya, East Java 60231, Indonesia.
Digital Object Idenfier (DOI): https://doi.org/10.35882/ijeeemi.v7i4.271

Although quantitative analysis demonstrates the superiority of our model, it is crucial to conduct a more in-depth qualitative evaluation of the generated report outputs. The primary challenge encountered in this comparison is the difficulty of matching the test data used in our study with specific datasets or samples from previous studies, as most prior works do not provide information regarding which images were used as illustrative results, nor has a standardized data split (such as the Karpathy split commonly adopted in general image captioning tasks) been established within this domain. Nevertheless, we successfully identified one image-caption pair that appeared in the CNX-B2 study [26], enabling a direct comparative analysis. The results of this comparison are presented in Fig. 14.

In Fig. 14, it can be observed that the CNX-B2 model generates reports that are structurally adequate but suffer from word truncation at the end of the sentence ("cardiomediastinal silhouette is un"), indicating potential issues with sequence fluency or prediction length control. In contrast, the report generated by our model exhibits a complete and coherent structure, with phrasing closely aligned with the ground truth, such as "the heart is normal in size" and "the lungs are clear," without omitting critical information such as "no focal consolidation, pneumothorax, or pleural effusion." Although CNX-B2 achieves slightly higher BLEU-2 and BLEU-3 scores (Table 6), our model demonstrates superiority in semantic completeness and visual description accuracy.

C. Research Limitations

Although the data balancing strategy through a combination of undersampling and oversampling successfully reduced bias toward the majority label, the model still had difficulty handling complex or very rare clinical cases due to dataset limitations that prevented representation of the full variety of medical findings. Consequently, the model tended to simplify descriptions or ignore specific clinical phrases that rarely appeared, indicating that the semantic representation of extreme minority findings was still not rich enough to support full generalization.

To better understand the model's failure modes, we conducted a qualitative error analysis on low-similarity cases (Fig. 10) and identified omission errors as a limitation. In several instances, the model failed to report key pathological findings explicitly stated in the ground truth. For example, in CXR3111, the reference report clearly describes "stable cardiomegaly" and "mild bilateral interstitial opacities… representing mild pulmonary edema," yet the generated report states "the heart is normal in size" and completely omits pulmonary edema. Such omissions are clinically significant, as they may lead to missed or delayed diagnoses, particularly in conditions such as heart failure, where accurate recognition of cardiomegaly and pulmonary congestion is essential for timely intervention.

A second prominent error type is hallucination, where the model introduces findings not present in the original report or supported by the image. As seen in CXR3390 and CXR2969, the model consistently generates phrases such as "emphysematous changes of the skeletal structures" and "atherosclerotic changes of the spine," despite their absence in the ground truth. These fabricated statements risk being misleading, potentially triggering unnecessary follow-up tests or incorrect assumptions about patient comorbidities.

These error types are especially concerning because they cannot be easily detected or corrected without insight into the model's reasoning process. Additionally, an important limitation of our current framework is the absence of built-in interpretability. The model can generate radiology reports but does not provide visual justifications for its clinical statements, making it difficult for radiologists to verify the accuracy of the findings. In high-risk medical contexts, such opacity may undermine trust and hinder clinical adoption. To address this, future versions should integrate explainability mechanisms that highlight the regions of the CXR image most influential in generating each sentence.

Although our combined undersampling and oversampling strategy effectively reduced label imbalance and improved overall metric scores, this approach involves trade-offs that require careful consideration. The undersampling of frequently occurring caption-projection pairs may inadvertently remove variations in common findings, potentially limiting the model's ability to distinguish subtle patterns. Meanwhile, despite applying image-space augmentations for minority labels, overly aggressive oversampling still poses a risk of overfitting, particularly for extremely rare conditions with limited visual diversity. Although the target frequency was capped at 250 to mitigate this risk, the possibility of overfitting remains. Future studies should explore the sensitivity of model performance to sampling thresholds to identify the optimal trade-off.

This limitation is exacerbated by the fact that the experiments did not involve hyperparameter tuning, as most parameters such as the learning rate and dropout rate were set based on common practice rather than optimization, leaving the model's potential for optimal performance unexplored. To address this in future work, and due to computational constraints, we plan to implement a sequential hyperparameter tuning strategy, optimizing one hyperparameter at a time (e.g., dropout rate or embedding dimension) while fixing others at their current best values. This approach enables efficient exploration of the hyperparameter space under limited GPU memory.

Furthermore, the absence of text augmentation techniques, such as the sentence reordering strategy proposed by [25], limited the model's exposure to diverse linguistic formulations of identical clinical findings. Specifically, [25] demonstrated that decomposing radiology reports into individual sentences and randomly shuffling their order during each training epoch can generate syntactically varied yet clinically consistent captions. This approach preserves diagnostic meaning while enriching structural diversity, thereby encouraging the model to learn pathology representations that are independent of fixed narrative patterns. If such a technique were integrated into our framework, it could

Corresponding author: Ricky Eka Putra, rickyeka@unesa.ac.id, Department of Informatics, State University of Surabaya, Jl. Ketintang, Ketintang, Kec. Gayungan, Surabaya, East Java 60231, Indonesia.
Digital Object Idenfier (DOI): https://doi.org/10.35882/ijeeemi.v7i4.271

enhance the model's robustness to inter-radiologist writing-style variations and reduce its tendency to produce rigid or repetitive phrasing.

The model proposed in this study has the potential to be a valuable tool for radiologists, especially in healthcare facilities with a shortage of skilled personnel or in remote areas with limited access to radiology services [5], [7], [8]. The system does not aim to replace medical professionals but rather to alleviate the initial workload by providing accurate and structured draft reports, allowing radiologists to focus on verification, refinement, and more complex clinical decision-making. Furthermore, implementing this automated system can improve consistency and standardization in radiology reporting, as the model generates output based on objectively studied data patterns rather than on individual preferences or writing styles. This can reduce variability among radiologists and improve the institutional quality of medical reporting [5].

The methodological implications of this study confirm that the success of a medical image captioning system depends on the synergy between model architecture complexity and a comprehensive data-handling approach. A data-balancing strategy combining undersampling and oversampling proved effective in reducing majority bias, while the integration of the CoAtNet hybrid architecture with a Bi-LSTM decoder and Multi-Head Attention improved output quality through bidirectional context understanding [22], [24]. These findings suggest that architectural innovation must be accompanied by careful data preprocessing to build accurate, fair, and clinically relevant systems, thereby enabling the development of better medical generative models.

Although quantitative metrics such as BLEU and ROUGE-L provide useful indications of linguistic similarity and structural coherence, they are insufficient for assessing clinical validity [18], [21]. High metric scores do not always guarantee diagnostic accuracy or clinical reliability, as generated reports may appear linguistically fluent yet still risk omitting critical findings or introducing pathologies unsupported by visual evidence. Therefore, the true value of an automated medical report generation system must be evaluated not only through automated metrics but also through expert clinical assessment [33].

As shown in Fig. 9, the model-generated reports demonstrate high textual alignment with expert radiologist reports, particularly in describing common or straightforward findings such as normal heart size, clear lungs, and the absence of pleural effusion. These examples illustrate the model's ability to capture essential clinical meaning with coherent sentence structure. However, as seen in Fig. 10, discrepancies remain in complex or rare cases, where the model tends to oversimplify descriptions or omit uncommon abnormalities. These qualitative aspects have already been discussed in detail in earlier subsections and collectively indicate that the model's clinical competence remains dependent on case complexity and linguistic variation, necessitating further evaluation under real-world diagnostic conditions.

In conclusion, the developed system is not intended to replace radiologists but rather to serve as an assistive drafting tool that produces structured initial report drafts for subsequent verification and refinement by qualified medical professionals. By automating the initial reporting stage, this model has the potential to reduce radiologists' workload and improve reporting consistency, particularly in healthcare settings with limited access to expert personnel [7], [8]. Future research should involve radiologists in clinical evaluations. Such expert assessment will be crucial to ensure the system's safety, reliability, and readiness for real-world clinical adoption [33].

V. Conclusion

We developed an automatic medical report generation system from CXR images by addressing data imbalance and limitations in contextual understanding found in previous models through a combination of undersampling strategies based on majority captions and oversampling based on minority labels, CoAtNet as a visual encoder, and Bi-LSTM with MHA for text processing. Experimental results show that the model achieved the highest cumulative score of 1.642, with B-1 (0.480), B-2 (0.329), B-3 (0.245), B-4 (0.183), and R-L (0.405), proving that the data balancing strategy reduces majority bias and that Bi-LSTM improves bidirectional contextual understanding to produce medical reports close to the ground truth.

Based on the identified limitations, future research is recommended to expand experiments using larger and more diverse datasets, such as MIMIC-CXR, which offers high data volume, or EGD-CXR, which provides a more balanced data distribution. Furthermore, systematic exploration of optimal sampling thresholds is needed to identify the best trade-off between bias reduction and preservation of clinical variation. To bridge the gap between algorithmic performance and real-world clinical utility, evaluations involving board-certified radiologists are essential. The application of hyperparameter tuning techniques is also necessary to optimally explore the parameter space and unlock the model's maximum performance potential. Finally, the integration of text augmentation techniques, such as syntactic reordering that preserves clinical meaning, should be explored, as these have proven effective in prior studies. This combined approach is expected to enhance the model's ability to capture more nuanced and clinically relevant findings.

References

[1] G. D. Ancona *et al.*, "Deep learning to predict long-term mortality from plain chest X-ray in patients referred for suspected coronary artery disease," *J Thorac Dis*, vol. 16, no. 8, pp. 4914–4923, Aug. 2024, doi: 10.21037/JTD-24-322/PRF.

[2] J. Mahawar and A. Paul, "Generalizable diagnosis of chest radiographs through attention-guided decomposition of images utilizing self-consistency loss," *Comput Biol Med*, vol. 180, p. 108922, Sep.

Corresponding author: Ricky Eka Putra, rickyeka@unesa.ac.id, Department of Informatics, State University of Surabaya, Jl. Ketintang, Ketintang, Kec. Gayungan, Surabaya, East Java 60231, Indonesia.
Digital Object Idenfier (DOI): https://doi.org/10.35882/ijeeemi.v7i4.271

2024, doi: 10.1016/J.COMPBIOMED.2024.108922.

[3] A. R. Alruwaili *et al.*, "A Critical Examination of Academic Hospital Practices—Paving the Way for Standardized Structured Reports in Neuroimaging," *Journal of Clinical Medicine 2024, Vol. 13, Page 4334*, vol. 13, no. 15, p. 4334, Jul. 2024, doi: 10.3390/JCM13154334.

[4] S. Harsini, S. Tofighi, L. Eibschutz, B. Quinn, and A. Gholamrezanezhad, "An Evolution of Reporting: Identifying the Missing Link," 2022. doi: 10.3390/diagnostics12071761.

[5] N. Kaur, A. Mittal, and G. Singh, "Methods for automatic generation of radiological reports of chest radiographs: a comprehensive survey," *Multimed Tools Appl*, vol. 81, no. 10, 2022, doi: 10.1007/s11042-021-11272-6.

[6] R. Riechie, V. Jessica, M. Kurniawan, and F. V. P. Samosir, "Convolutional Kolmogorov-Arnold Network for Pneumonia Detection in Medical Image Analysis," *Indonesian Journal of Electronics, Electromedical Engineering, and Medical Informatics*, vol. 7, no. 3, pp. 475–487, Aug. 2025, doi: 10.35882/IJEEEMI.V7I3.106.

[7] C. L. Canon, J. F. B. Chick, I. DeQuesada, R. B. Gunderman, N. Hoven, and A. E. Prosper, "Physician Burnout in Radiology: Perspectives From the Field," *American Journal of Roentgenology*, vol. 218, no. 2, 2022, doi: 10.2214/AJR.21.26756.

[8] C. H. Ko, L. N. Chien, Y. T. Chiu, H. H. Hsu, H. F. Wong, and W. P. Chan, "Demands for medical imaging and workforce Size: A nationwide population-based Study, 2000–2020," *Eur J Radiol*, vol. 172, 2024, doi: 10.1016/j.ejrad.2024.111330.

[9] 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.

[10] E. J. Topol, "Toward the eradication of medical diagnostic errors," 2024. doi: 10.1126/science.adn9602.

[11] Z. Ye, R. Khan, N. Naqvi, and M. S. Islam, "A novel automatic image caption generation using bidirectional long-short term memory framework," *Multimed Tools Appl*, vol. 80, no. 17, 2021, doi: 10.1007/s11042-021-10632-6.

[12] Y. Ming, N. Hu, C. Fan, F. Feng, J. Zhou, and H. Yu, "Visuals to Text: A Comprehensive Review on Automatic Image Captioning," *IEEE/CAA Journal of Automatica Sinica*, vol. 9, no. 8, 2022, doi: 10.1109/JAS.2022.105734.

[13] 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.

[14] A. Ueda, W. Yang, and K. Sugiura, "Switching Text-Based Image Encoders for Captioning Images With Text," *IEEE Access*, vol. 11, 2023, doi: 10.1109/ACCESS.2023.3282444.

[15] 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.

[16] R. Sasibhooshan, S. Kumaraswamy, and S. Sasidharan, "Image caption generation using Visual Attention Prediction and Contextual Spatial Relation Extraction," *J Big Data*, vol. 10, no. 1, 2023, doi: 10.1186/s40537-023-00693-9.

[17] 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.

[18] 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.

[19] D. Hou, Z. Zhao, Y. Liu, F. Chang, and S. Hu, "Automatic Report Generation for Chest X-Ray Images via Adversarial Reinforcement Learning," *IEEE Access*, vol. 9, 2021, doi: 10.1109/ACCESS.2021.3056175.

[20] Y. Gu, R. Li, X. Wang, and Z. Zhou, "Automatic Medical Report Generation Based on Cross-View Attention and Visual-Semantic Long Short Term Memorys," *Bioengineering*, vol. 10, no. 8, 2023, doi: 10.3390/bioengineering10080966.

[21] 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.

[22] 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.

[23] D. Naik and C. D. Jaidhar, "A novel Multi-Layer Attention Framework for visual description prediction using bidirectional LSTM," *J Big Data*, vol. 9, no. 1, 2022, doi: 10.1186/s40537-022-00664-6.

[24] H. Zhang, C. Ma, Z. Jiang, and J. Lian, "Image Caption Generation Using Contextual Information Fusion With Bi-LSTM-s," *IEEE Access*, vol. 11, 2023, doi: 10.1109/ACCESS.2022.3232508.

Corresponding author: Ricky Eka Putra, rickyeka@unesa.ac.id, Department of Informatics, State University of Surabaya, Jl. Ketintang, Ketintang, Kec. Gayungan, Surabaya, East Java 60231, Indonesia.
Digital Object Idenfier (DOI): https://doi.org/10.35882/ijeeemi.v7i4.271

[25] D. Parres, A. Albiol, and R. Paredes, "Improving Radiology Report Generation Quality and Diversity through Reinforcement Learning and Text Augmentation," *Bioengineering 2024, Vol. 11, Page 351*, vol. 11, no. 4, p. 351, Apr. 2024, doi: 10.3390/BIOENGINEERING11040351.

[26] 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.

[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.

[28] 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.

[29] Z. Dai, H. Liu, Q. V. Le, and M. Tan, "CoAtNet: Marrying Convolution and Attention for All Data Sizes," in *Advances in Neural Information Processing Systems*, 2021.

[30] S. M. N. Ashraf, M. A. Mamun, H. M. Abdullah, and M. G. R. Alam, "SynthEnsemble: A Fusion of CNN, Vision Transformer, and Hybrid Models for Multi-Label Chest X-Ray Classification," in *2023 26th International Conference on Computer and Information Technology, ICCIT 2023*, 2023. doi: 10.1109/ICCIT60459.2023.10441433.

[31] T. Xie, W. Ding, J. Zhang, X. Wan, and J. Wang, "Bi-LS-AttM: A Bidirectional LSTM and Attention Mechanism Model for Improving Image Captioning," *Applied Sciences (Switzerland)*, vol. 13, no. 13, Jul. 2023, doi: 10.3390/app13137916.

[32] 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.

[33] P. Sloan, P. Clatworthy, E. Simpson, and M. Mirmehdi, "Automated Radiology Report Generation: A Review of Recent Advances," *IEEE Rev Biomed Eng*, vol. 18, pp. 368–387, 2025, doi: 10.1109/RBME.2024.3408456.

[34] G. Siracusano, A. La Corte, A. G. Nucera, M. Gaeta, M. Chiappini, and G. Finocchio, "Effective processing pipeline PACE 2.0 for enhancing chest x-ray contrast and diagnostic interpretability," *Sci Rep*, vol. 13, no. 1, 2023, doi: 10.1038/s41598-023-49534-y.

[35] H. Deng, H. Zhao, H. Zhang, and G. Liu, "γ Radiation Image Enhancement Method Based on Non-Linear Mapping," *IEEE Access*, vol. 10, 2022, doi: 10.1109/ACCESS.2022.3209807.

[36] R. Hafizah, T. H. Saragih, M. Muliadi, F. Indriani, and M. I. Mazdadi, "Machine Learning Implementation for Sentiment Analysis on X/Twitter: Case Study of Class Of Champions Event in Indonesia," *Indonesian Journal of Electronics, Electromedical Engineering, and Medical Informatics*, vol. 7, no. 2, pp. 370–386, May 2025, doi: 10.35882/IJEEEMI.V7I2.81.

[37] M. S. Alam *et al.*, "Attention-based multi-residual network for lung segmentation in diseased lungs with custom data augmentation," *Sci Rep*, vol. 14, no. 1, pp. 1–11, Dec. 2024, doi: 10.1038/S41598-024-79494-W;SUBJMETA=114,631,692,698;KWRD=ANATOMY,COMPUTATIONAL+BIOLOGY+AND+BIOINFORMATICS.

[38] Z. Fei, M. Fan, L. Zhu, J. Huang, X. Wei, and X. Wei, "Uncertainty-Aware Image Captioning," in *Proceedings of the 37th AAAI Conference on Artificial Intelligence, AAAI 2023*, 2023. doi: 10.1609/aaai.v37i1.25137.

[39] H. Huan, J. Yan, Y. Xie, Y. Chen, P. Li, and R. Zhu, "Feature-enhanced nonequilibrium bidirectional long short-term memory model for Chinese text classification," *IEEE Access*, vol. 8, 2020, doi: 10.1109/ACCESS.2020.3035669.

[40] M. Sirshar, M. F. K. Paracha, M. U. Akram, N. S. Alghamdi, S. Z. Y. Zaidi, and T. Fatima, "Attention based automated radiology report generation using CNN and LSTM," *PLoS One*, vol. 17, no. 1 January, 2022, doi: 10.1371/journal.pone.0262209.

[41] 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.

[42] J. Wang, W. Xu, Q. Wang, and A. B. Chan, "On Distinctive Image Captioning via Comparing and Reweighting," *IEEE Trans Pattern Anal Mach Intell*, vol. 45, no. 2, 2023, doi: 10.1109/TPAMI.2022.3159811.

[43] J. Chen, "Transform, contrast and tell: Coherent entity-aware multi-image captioning," *Computer Vision and Image Understanding*, vol. 238, p. 103878, Jan. 2024, doi: 10.1016/J.CVIU.2023.103878.

[44] Z. Liu *et al.*, "Swin Transformer V2: Scaling Up Capacity and Resolution," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2022. doi: 10.1109/CVPR52688.2022.01170.

[45] S. Woo *et al.*, "ConvNeXt V2: Co-designing and Scaling ConvNets with Masked Autoencoders," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern*

Corresponding author: Ricky Eka Putra, rickyeka@unesa.ac.id, Department of Informatics, State University of Surabaya, Jl. Ketintang, Ketintang, Kec. Gayungan, Surabaya, East Java 60231, Indonesia.
Digital Object Idenfier (DOI): https://doi.org/10.35882/ijeeemi.v7i4.271

Recognition, 2023. doi: 10.1109/CVPR52729.2023.01548.

[46] 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.

AUTHOR BIOGRAPHY

Rafy Aulia Akbar received the Bachelor's degree in Informatics Engineering from Universitas Negeri Surabaya, Indonesia, in 2019, with a concentration in Computer Science. He is currently pursuing his Master's degree in Informatics at the same university, focusing on Computer Science with research interests in Computer Vision, Medical Image Analysis, Natural Language Processing, and Deep Learning. During his undergraduate studies, he received the "Best Solution Award" at the ASEAN Vocational and Engineering Camp 2017 "Toward School 4.0," a regional competition participated in by teams from multiple ASEAN countries. His current academic work explores the intersection of multimodal deep learning, particularly in automated medical report generation from chest X-ray images.

Ricky Eka Putra received the Bachelor's degree in Computer Science from Institut Teknologi Sepuluh Nopember (ITS), Surabaya, Indonesia, in 2008, followed by the Master's degree in Computer Science in 2011and the Doctoral degree in 2020, all from the same institution. He obtained his Professional Engineer title in 2024. He currently serves as a Lecturer in the Department of Informatics at Universitas Negeri Surabaya. His research interests include Deep Learning, Big Data Analytics, Digital Applications and User Experience, Clinical Psychology and Behavioral Intention, Healthcare Informatics, Statistical Methods, and Digital and Early Education. He is particularly focused on developing intelligent systems that bridge technology and clinical needs.

Wiyli Yustanti received the Bachelor's degree in Science from Institut Teknologi Sepuluh Nopember (ITS), Surabaya, Indonesia, in 1999, followed by the Master's degree in 2004 and the Doctoral degree in 2024 from the same institution. She is currently a Senior Lecturer in the Department of Informatics at Universitas Negeri Surabaya. Her research interests include Deep Learning, Unsupervised Learning, Sentiment Analysis, Computer Vision, Cloud and Data Security, Clinical Psychology and Behavioral Data, and Big Data Analytics. Her recent work explores the integration of computer vision and natural language processing for multimodal data analysis, aiming to support early detection systems and personalized interventions in mental and physical health domains.

Corresponding author: Ricky Eka Putra, rickyeka@unesa.ac.id, Department of Informatics, State University of Surabaya, Jl. Ketintang, Ketintang, Kec. Gayungan, Surabaya, East Java 60231, Indonesia.
Digital Object Idenfier (DOI): https://doi.org/10.35882/ijeeemi.v7i4.271