Automated radiology report generation from chest X-ray images using CheXNet and Transformer-LSTM architecture

Mohamed Adel[1,†], Mohamed Aborizka[2,*,†]

[1] *Arab Academy for Science and Technology, Sheraton, 11757, Cairo, Egypt*

Abstract

Automated radiology report generation from chest X-ray (CXR) images has the potential to significantly reduce the workload of radiologists and improve diagnostic efficiency. In this paper, we propose a novel architecture that integrates a Transformer encoder with an LSTM layer to generate coherent, contextually accurate reports from CXR images. CheXNet, a DenseNet-based model pre-trained on the Chest X-ray dataset, is employed to extract 1024-dimensional feature vectors from the input images. These features are passed to the Transformer encoder, which uses multi-head attention and positional encodings to capture both local and global relationships in the data. An LSTM layer is introduced after the encoder to refine the image features and better capture sequential dependencies in the report. The Transformer decoder generates the report in an autoregressive manner, utilizing beam search during inference to improve fluency and accuracy. Experimental results show that our model achieves competitive performance across BLEU-1 to BLEU-4 scores, with a BLEU-1 score of 0.4636 and a BLEU-4 score of 0.3575, outperforming several baseline methods. The results indicate that our hybrid approach effectively balances word-level accuracy and sequence coherence, making it a robust solution for medical report generation.

Keywords

Natural Language Processing (NLP), Transformer, Beam Search, ChexNet, Long Short-Term Memory (LSTM), Bilingual evaluation understudy (BLEU)

1. Introduction

Automated medical report generation has gained increasing attention in recent years, driven by the growing volume of medical imaging data and the limited availability of radiologists. Chest X-ray (CXR) images, one of the most commonly used diagnostic tools in healthcare, provide critical insights into thoracic conditions, including pneumonia, lung cancer, and cardiovascular diseases [1] . Despite their diagnostic significance, interpreting these images and drafting comprehensive radiology reports remain time-consuming and require specialized expertise. Automating the generation of medical reports from CXR images could significantly reduce the workload for radiologists, accelerate diagnosis, and improve the consistency of reporting [2]. However, generating coherent, clinically accurate radiology reports is a challenging task. Unlike general image captioning, which focuses on describing visual content in natural language, radiology report generation requires a deeper understanding of both visual features and medical domain-specific language [3].These reports not only describe visual abnormalities but also provide diagnostic conclusions [4] , making it crucial for models to capture both global and local features of the image and produce contextually relevant sequences that follow medical conventions. Moreover, generating accurate medical reports requires the model to integrate multimodal information, combining visual data from the images with linguistic structures in medical terminology. The complexity of medical language, which often includes abbreviations, specialized terms, and implicit contextual knowledge, adds another layer of difficulty. Additionally, reports need to be not only factually accurate but also aligned with clinical standards, which makes it necessary for the model to be highly reliable in real-world settings. Current research is focused on developing more robust and interpretable models that can handle these multifaceted challenges.

IDDM'24: 7th International Conference on Informatics & Data-Driven Medicine, November 14 - 16, 2024, Birmingham, UK
* *Corresponding author.* † *These authors contributed equally.*
mohamed_adel_98@yahoo.com (M. Adel); m.aborizka@aast.edu (M. Aborizka)
0009-0006-0315-5333 (M. Adel); 0000-0003-1154-6407 (M. Aborizka)

Several studies have explored automated report generation using various deep learning techniques. For instance, CheXNet, a DenseNet-based model pre-trained on the Chest X-ray dataset, has shown promising results in detecting thoracic diseases from CXR images [2] . Other approaches have incorporated convolutional neural networks (CNNs) to extract image features, paired with recurrent neural networks (RNNs) or long short-term memory (LSTM) networks to generate reports [5]. While these methods demonstrate the potential for automated report generation, they often struggle with producing coherent and contextually rich sentences, particularly when generating longer sequences of text.

In this study, we propose a novel architecture that integrates a Transformer-based model with an LSTM layer to address these challenges. Transformers, known for their ability to capture long-range dependencies through attention mechanisms [6], have revolutionized natural language processing (NLP) tasks. By incorporating multi-head attention and positional encodings, Transformers can effectively model complex relationships in visual data. However, their lack of inherent sequential modeling capabilities can limit their ability to generate text that follows a logical progression. To overcome this limitation, we augment the Transformer encoder with an LSTM layer, which is well-suited for capturing temporal dependencies and sequential information. This hybrid architecture allows us to model both the spatial features of CXR images and the sequential nature of medical reports. Furthermore, our model employs CheXNet for feature extraction, leveraging its pre-trained knowledge on thoracic disease detection to enhance the accuracy of the generated reports [5]. By combining CheXNet's feature extraction capabilities with the powerful attention mechanisms of the Transformer and the sequential modeling strengths of LSTM, our approach aims to produce more accurate and contextually coherent radiology reports.

Main Contributions:

1. **Hybrid Transformer-LSTM Architecture:** We propose a novel model that combines a Transformer encoder with an LSTM layer to address the challenges of generating coherent, sequential radiology reports from CXR images. The model utilizes both self-attention and cross-attention mechanisms, self-attention in the encoder to capture global dependencies within the image features, and cross-attention in the decoder to align the generated text with the encoded image features. This architecture leverages the strengths of both attention mechanisms and sequential modeling to improve the quality of report generation.
2. **CheXNet Feature Extraction:** We integrate CheXNet, a DenseNet-based model pre-trained on the Chest X-ray dataset, to extract rich 1024-dimensional feature vectors from CXR images. CheXNet's ability to detect thoracic diseases enhances the visual representation passed to the Transformer encoder.
3. **Beam Search Decoding:** During the inference phase, we implement beam search with varying beam widths (2, 5, and 7) to improve the fluency and accuracy of the generated reports. Beam search allows the model to explore multiple word sequences, enhancing the quality of the final output.
4. **Comprehensive Evaluation:** We evaluate the model's performance using BLEU scores across multiple n-gram levels (BLEU-1, BLEU-2, BLEU-3, and BLEU-4), demonstrating its effectiveness in capturing both word-level accuracy and contextual coherence. Additionally, we compare the performance of our approach with several existing methods in the field, highlighting its advantages and areas for further improvement.

By addressing both the spatial and sequential aspects of report generation, this work presents a robust framework for automating radiology report generation from CXR images. Figure 1 illustrates the architecture of the proposed system and the flow of data throughout the model's processing stages. The paper at hand comprises the following: Section II presents the related work. Section III presents the material and methods of the work. Experiments are illustrated in Section IV. Section V presents the experimented results. Section VI presents the discussion. Finally, section VII concludes the paper.

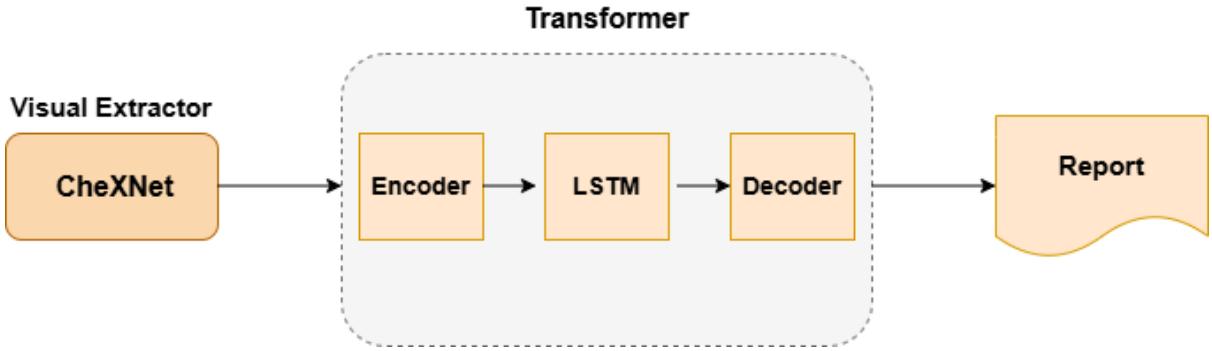

Figure 1: Proposed System

2. Related Work

A significant amount of research work has been conducted on automatic report generation from radiological images. Li et al. [7] propose a model that leverages disease graphs for medical report generation. Their approach involves transforming visual features into a structured abnormality graph using an encoding module. Liu et al. [8] introduced a domain-aware automatic report generation model specifically for chest radiography. They incorporated reinforcement learning to refine the readability and clinical accuracy of the generated reports. Zhang et al. [9] designed a module that integrates a pre-built graph of chest abnormalities across various diseases to enhance report generation. Their model uses attention mechanisms and graph convolution to learn embedded features from the graph. Lovelace et al. [10] presented a transformer-based neural machine translation model that fine-tunes clinical data extraction from reports, improving consistency and clinical relevance. Lastly, Chen et al. [11] developed a memory-driven transformer model for generating X-ray reports, which demonstrated superior performance in both language generation metrics and clinical assessment compared to previous models. Amjoud et al. [12] propose a deep learning model for automatic CXR report generation, combining a modified transformer architecture with pre-trained CNNs for feature extraction. Their model addresses the complexity of medical report generation, which requires domain-specific language and coherent, accurate content. The proposed system includes three sub-models: a pre-trained CNN for feature extraction, and an encoder-decoder transformer for report generation. They evaluated their model on the OpenI Indiana University CXR dataset, consisting of 2955 reports and 6091 images after preprocessing. Using BLEU, METEOR, and ROUGE metrics, the model was assessed for its performance. Elaanba et al. [13] propose a transformer-based model for generating radiology text reports from CXR images using both frontal and lateral views. Their model leverages the Vision Transformer (ViT) as a feature extractor, which outperforms traditional CNN-based models like DenseNet-121. They also explore the effect of using dual-view input (frontal and lateral images) versus single-view input, with dual-view input leading to better performance in generating accurate reports. The experiments were conducted on the IU-X-ray dataset, and the model's quality was evaluated using the BLEU metric. The study found that dual-view input improves report generation accuracy, achieving a BLEU-1 score of 0.29, while models using single-view inputs scored lower. The paper highlights the importance of high-quality data annotation and suggests further experiments on dual-view inputs. Jing et al. [5] propose a multi-task learning framework for generating medical imaging reports, addressing key challenges such as localizing abnormalities and generating long, coherent descriptions. Their model incorporates a co-attention mechanism to align visual and textual features, and a hierarchical LSTM network to effectively produce detailed reports.

3. Materials and Methods

3.1. CheXNet

Convolutional neural networks (CNNs) are a class of artificial neural networks that are widely used for feature extraction and classification tasks, particularly in image and time-series data [14]. CheXNet is a specialized CNN-based deep learning model designed to analyze CXR images and detect various diseases. It is built on the DenseNet-121 architecture, which is known for its efficiency and effectiveness in image classification tasks [15]. The pre-trained weights of CheXNet enhance the model's ability to identify thoracic diseases from CXR images, reducing the need for extensive re-training and improving feature extraction accuracy. The architecture of CheXNet, specifically DenseNet-121, incorporates multiple dense blocks and transition layers. Dense blocks consist of several convolutional layers, where the output of each layer is concatenated with the outputs from previous layers [2]. Within each dense block, convolutional layers typically use a combination of 1x1 and 3x3 convolutions. The 1x1 convolutions, also known as bottleneck layers, are used to reduce the dimensionality of the input feature maps, which decreases computational complexity and accelerates the learning process [16]. The 3x3 convolutions then capture more complex features by analyzing local regions of the image. By stacking these convolutional layers, the model can learn intricate patterns and representations from the input images. This dense connectivity enables the network to learn a rich set of features, with each layer building upon the features extracted by earlier layers [15]. Between dense blocks, CheXNet incorporates transition layers that perform down-sampling. These transition layers consist of convolutional operations followed by pooling layers, which reduce the spatial dimensions of the feature maps. This down-sampling helps in managing computational resources and controlling overfitting by reducing the number of feature maps and the spatial dimensions of the data. CheXNet's application in feature extraction is particularly valuable due to its ability to produce a rich set of hierarchical features from CXR images. In feature extraction, the model identifies and captures important patterns and characteristics from images. The dense connectivity and deep structure of CheXNet allow it to capture a range of features, from simple patterns such as edges and textures in lower layers to more complex and disease-related features in higher layers [17]. This hierarchical feature extraction is crucial for detecting subtle and complex patterns in medical images. Utilizing CheXNet for feature extraction is advantageous because the model's comprehensive feature set aids in diagnosing medical conditions. The ability to capture detailed and nuanced features enhances diagnostic accuracy and provides valuable insights into the presence and severity of various diseases. These extracted features can also be used in conjunction with other machine learning models or algorithms to further improve diagnostic performance or to analyze patterns and correlations within the dataset.

3.2. Encoder-Decoder Transformer

A transformer based encoder-decoder architecture is used to generate natural language reports from CXR images. The encoder-decoder transformer is an advanced model designed for sequence-to-sequence tasks, particularly useful for transforming image data into coherent textual reports. Below, the key components of the model are described, focusing on how the encoder and decoder work together to accomplish this task.

3.2.1. Encoder

The encoder's role is to encode the visual information from CXR images and convert it into a set of encoded features. Unlike traditional transformer models that rely solely on self-attention [6], this design uses self-attention to process visual features extracted from the image [18], enabling the model to identify relationships within the image data. The input to the encoder consists of feature maps produced by a DenseNet model, which acts as the visual feature extractor [15]. DenseNet processes the X-ray image and generates feature vectors that are then fed into the encoder. The encoder transforms these input features, denoted as xi, into a sequence of hidden states, hi, which represent high-level, compressed features of the image. These hidden states capture essential image information required for generating the report. Each hidden state hi, corresponds to a portion of the image's content and plays a crucial role in guiding the decoder to produce relevant words in the

report. In contrast, the decoder incorporates cross-attention, which attends to the encoded image features while generating the report. The cross-attention mechanism enables the decoder to consider both the image features and the previously generated words, ensuring the output is both visually accurate and contextually coherent. This allows the model to integrate visual patterns and contextual cues from the partially generated report during decoding. Mathematically, the transformation process in the encoder can be described by the self-attention mechanism, which focuses purely on the visual features.

$$hi = SelfAttention(xi) \tag{1}$$

here xi represents the visual features extracted by the DenseNet from the X-ray image, and hi denotes the encoded hidden state for each segment of the input feature map.

3.2.2. Decoder

The decoder generates a natural language report based on the encoded features from the encoder. It produces a sequence of words, with each word depending on both the encoded image features and the previously generated words in the report. To improve the decoding process, a relational memory module is integrated into the transformer architecture [19] . This, along with memory-driven conditional layer normalization (MCLN), helps the decoder maintain context throughout the report, ensuring consistency and relevance during sentence generation. The decoding process uses self-attention to ensure coherence by attending to previously generated words and cross-attention to the encoded image features to ensure the generated text reflects the visual information accurately [6]. The cross-attention mechanism calculates a weighted sum of the encoded features, producing a context vector that contains the most relevant image information. This context vector is essential for guiding the generation of the next word. The decoder then processes the context vector through a feedforward neural network and applies a softmax activation function to generate a probability distribution over possible words in the output vocabulary.

The report generation can be formalized using the following chain rule, where the probability of generating the next word depends on both the prior words and the encoded image features:

$$P(y_t|y_1, y_2, \dots, y_{t-1}, Img; \theta) \tag{2}$$

where $\boldsymbol{Y}$ is the target text sequence (the report), $\boldsymbol{Img}$ refers to the input X-ray image, and $\boldsymbol{\theta}$ represents the model parameters. During training, the goal is to maximize the log-likelihood of generating the correct sequence $\boldsymbol{Y}$, given the input image. At inference time, the model uses beam search to iteratively sample words from the probability distribution, generating the report one word at a time until it reaches a predefined maximum length or an end-of-sentence token [18]. In summary, this encoder-decoder transformer model extracts detailed visual features from CXR images and converts them into accurate and coherent radiology reports. By utilizing self-attention in the encoder to process image features and cross-attention in the decoder to merge visual and textual information, the model ensures that each generated word is both medically relevant and linguistically coherent [19].

3.3. Beam Search

Beam Search is a widely used decoding algorithm for generating sequences in tasks such as machine translation, image captioning, and text generation [20]. It is an extension of the greedy search algorithm, where instead of selecting only the highest probability token at each time step, beam search maintains a set of the top $\boldsymbol{k}$ candidate sequences, referred to as the beam width. This allows

the algorithm to explore multiple possible sequences concurrently, balancing local optimality and global coherence in the generated output. At the initial step, the model begins with the start token ***startseq*** and generates a probability distribution over the vocabulary for the next word. Rather than selecting only the word with the highest probability, beam search selects the top $\boldsymbol{k}$ words with the highest probabilities. These $\boldsymbol{k}$ words form the beginning of $\boldsymbol{k}$ different candidate sequences. The likelihood of a sequence $\boldsymbol{S} = \{\boldsymbol{w_1}, \boldsymbol{w_2}, \dots \dots, \boldsymbol{w_n}\}$ is given by the product of the probabilities of each word in the sequence:

$$P(S) = \prod_{i=1}^{n} P(w_i | w_1, w_2, \dots \dots, w_{i-1}) \tag{3}$$

However, since multiplying probabilities can lead to very small numbers, beam search typically operates on the logarithmic scale to avoid underflow and to convert the product into a sum:

$$\log P(S) = \sum_{i=1}^{n} \log P(w_i | w_1, w_2, \dots \dots, w_{i-1})) \tag{4}$$

At each subsequent time step, the algorithm expands each of the current $\boldsymbol{k}$ sequences by appending the next word, again considering the top $\boldsymbol{k}$ words based on their conditional probabilities. The result is $\boldsymbol{k} \times \boldsymbol{k}$ times $\boldsymbol{k} \times \boldsymbol{k}$ candidate sequences, which are then pruned by selecting only the top $\boldsymbol{k}$ sequences with the highest cumulative scores. This process is repeated until one of the termination conditions is met, such as reaching a predefined maximum sequence length or encountering the end-of-sequence token ***endseq***. The beam width $\boldsymbol{k}$ plays a crucial role in controlling the trade-off between search depth and computational efficiency. A larger beam width allows the algorithm to explore a wider range of possible sequences, potentially leading to more accurate and coherent outputs. However, increasing $\boldsymbol{k}$ also results in higher computational costs, as the number of sequences that need to be evaluated and ranked grows exponentially [20]. Conversely, a smaller beam width reduces computational overhead but risks missing the globally optimal sequence by pruning too aggressively. To avoid biasing the search towards shorter sequences, the total score for each sequence is normalized by the sequence length:

$$Normalized\ score = \frac{score}{sequence\ length} \tag{5}$$

This normalization ensures that longer, more informative sequences are not penalized simply due to their length. Once the decoding process is complete, the sequence with the best (lowest) score is selected as the final output. Beam search is commonly applied to generate radiology reports from CXR images. Each image was processed to extract features, which were then fed into the captioning model to generate descriptive reports. Different beam widths can be explored to evaluate trade-offs between sequence diversity and computational complexity. The generated sequences are often evaluated using BLEU scores to measure the similarity between the predicted reports and the actual medical reports [21].

3.4. Long Short-Term Memory

The Long Short-Term Memory (LSTM) layer is a type of recurrent neural network (RNN) designed to capture patterns over long sequences of data, particularly useful in tasks where maintaining a logical sequence is essential. Unlike standard RNNs, LSTMs can retain important information over extended periods, making them ideal for applications where the order of information is crucial, such as text generation or time-series forecasting. By capturing sequential dependencies, LSTMs help ensure that each generated word or phrase aligns with prior context, producing coherent and contextually accurate outputs.

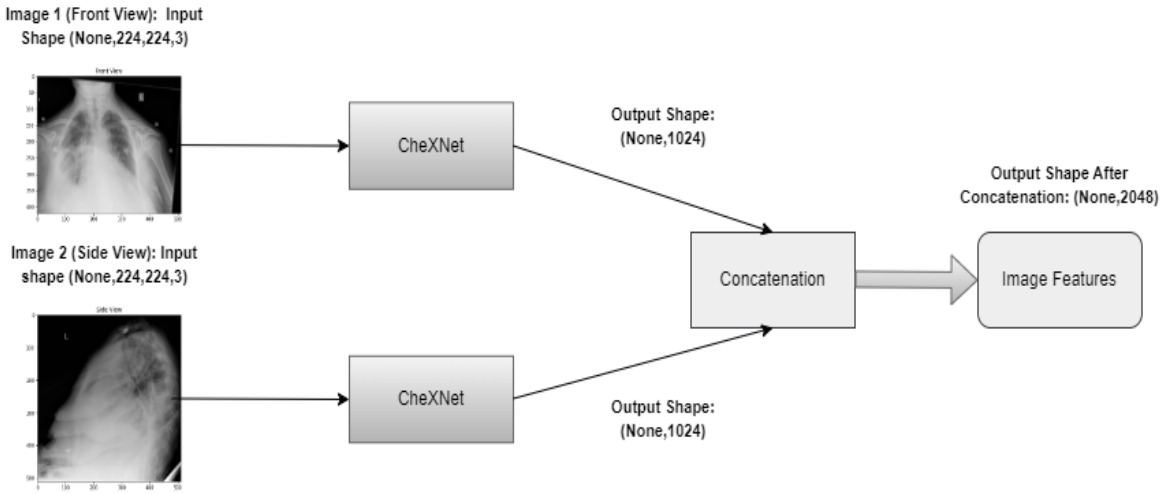

Figure 2: Multi-Input Feature Extraction Architecture

4. Experiments

4.1. Dataset

In this study, we utilized the public OpenI dataset, which contains CXR images collected from the Indiana University hospital network [22]. This dataset consists of two main components: a folder of X-ray images and a corresponding folder of radiography reports in XML format. Each radiography report can be linked to multiple images. Upon closer inspection, we observed that 2500 reports have two associated images, approximately 490 reports have only one image, 100 reports have three images, and 10 reports are associated with four images. To standardize the dataset, we focused on maintaining a consistent pairing of two images per report, selecting one frontal and one lateral view when applicable. For cases where reports had only one image, the image was duplicated to meet this standard. For reports that included four images, two side views and two front views, we selected one side view and one front view for inclusion in the dataset. This approach allowed for uniformity in data preparation, ensuring that each report would have exactly two associated images for the model's input. The final dataset was split into three subsets for training, validation, and testing purposes. Specifically, 2500 reports were allocated for training, 500 for validation, and 350 for testing [23].

4.2. Preprocessing

The dataset comprises CXR images in PNG format and corresponding radiology reports in XML format. To extract the relevant information, we first parsed the XML files using the ElementTree library [24] to retrieve sections such as comparison, indication, findings, and impressions, which are key components of radiology reports. This extraction was necessary to match the radiological findings with the images for downstream tasks. For each image, we also gathered metadata, such as image dimensions, by using the OpenCV library [25], which enabled efficient reading and

processing of image files. Handling missing values was a critical step to ensure a complete dataset for modeling. Many radiology reports had missing sections, such as findings or impressions. We addressed this by imputing default text such as "no findings" or "no impressions" where applicable [26]. This imputation was crucial to prevent empty fields from disrupting the training process, particularly for NLP tasks [27]. Moreover, for patients with missing images, we logged and skipped those entries to ensure image and text pairs were consistently maintained. Text preprocessing was applied to standardize the radiology reports for subsequent NLP analysis [27]. We converted all text to lowercase to reduce the complexity of the vocabulary and expanded common contractions to maintain linguistic clarity. Punctuation was removed to eliminate noise, except for full stops to preserve sentence boundaries, which are important in medical text for distinguishing between different clinical findings.

Additionally, we removed numerical values and sequences of irrelevant characters, such as repeated "x" marks, as they do not provide meaningful information for text analysis. Words with fewer than two characters were filtered out, except for medically relevant terms, ensuring that only informative words remained in the processed text. Finally, for patients with multiple X-ray images, we generated image pairs to capture the relationship between different views of the same anatomical region, such as frontal and lateral chest views. When multiple images were available for a single patient, we carefully paired them to ensure that both images corresponded to the same time point or diagnostic session, maintaining consistency in the dataset. In cases where only one image was available for a patient, we duplicated the image to create a pair, treating both images as the same view. This was done to standardize the input format for the model, which requires two images per patient for proper training. The rationale behind creating image pairs lies in the need for the model to learn the spatial and anatomical correlations between different views of the same region. In medical imaging, different angles can reveal additional information or confirm findings from another view, so combining these perspectives helps the model develop a more comprehensive understanding of the patient's condition. By exposing the model to multiple views of the same anatomical region during training, we enhance its ability to make more accurate predictions, particularly when subtle pathological findings might appear differently depending on the orientation. For each image pair, we ensured that the associated clinical findings from the radiology report were properly aligned with the visual data. This step was crucial in maintaining the integrity of the dataset, as accurate alignment between image pairs and their corresponding text annotations is vital for the success of multi-modal deep learning models. These reports provided context for the visual data, allowing the model to map specific radiological findings to certain visual patterns in the X-ray images, improving its interpretative capabilities.

This preprocessing pipeline of creating image pairs and aligning them with their corresponding textual reports facilitated the construction of a highly structured and coherent dataset. Figure 2 illustrates our proposed flowchart, outlining the preprocessing pipeline. This dataset was instrumental in the subsequent training, validation, and testing phases of the study, enabling the model to learn not only from the visual features of individual images but also from the relationships between different views and the corresponding medical context.

The resulting model was able to provide more accurate and nuanced diagnostic predictions, contributing to the overall goal of generating reliable radiology reports from CXRs. Finally, for patients with multiple images, we created image pairs by combining available images or using the same image for both positions when only one image was present. This step was essential for our model to understand the relationship between different views of the same anatomical region. Each image pair was associated with the corresponding findings from the report, ensuring proper alignment of textual and visual data. This preprocessing pipeline enabled the creation of a structured dataset that could be reliably used for the training, validation, and testing phases of the study.

4.3. Feature Extraction

In our research, the feature extraction process is fundamentally supported by the CheXNet model, which is built on the DenseNet-121 architecture [15]. DenseNet-121, an advanced variant of Dense Convolutional Networks, is characterized by its dense connectivity pattern, wherein each layer

receives input from all preceding layers. This design facilitates efficient gradient flow and feature propagation, addressing the vanishing gradient problem and enhancing feature reuse [15]. Originally pre-trained on the ImageNet dataset, DenseNet-121 has acquired the capability to extract general features from a diverse set of images. Fine-tuning CheXNet on the Chest X-ray dataset further specializes the model in identifying features relevant to CXR images, such as specific textures and patterns indicative of thoracic abnormalities [2]. To adapt CheXNet for feature extraction purposes, we removed the final fully connected (dense) layer, which is designed for classification tasks. This layer, specific to the classification of various diseases, was excluded to ensure that the output consists solely of the high-level feature maps produced by the convolutional layers [2]. These feature maps retain detailed spatial and semantic information about the CXR images, which is crucial for generating radiology reports.

Before feeding the images into the model, several preprocessing steps were applied. The images were resized to 224x224 pixels to match the input size expected by DenseNet-121, ensuring compatibility and consistency across the dataset. Additionally, grayscale CXR images were converted to RGB format by replicating the single grayscale channel across three channels, aligning with the model's requirement for RGB input [2]. Normalization was also performed by dividing pixel values by 255, which scales the values to the [0,1] range. This step standardizes the input data, stabilizing the training process, and improving convergence [2]. Once the images were pre-processed, they were passed through the modified CheXNet model to extract feature maps from the final convolutional layer [2]. These convolutional layers capture hierarchical features, ranging from low-level textures and edges to complex structures and patterns [15]. The resulting feature maps are high-dimensional representations of the CXR images, containing essential information about their spatial and textual characteristics. To convert these high-dimensional feature maps into a more compact form, we applied Global Average Pooling (GAP) [15]. GAP computes the average of all spatial locations within each feature map, resulting in a single scalar value per map. This dimensionality reduction helps prevent overfitting and reduces computational complexity, making the feature vectors more manageable for downstream tasks [15]. The output of the feature extraction process is a 1024-dimensional vector for each CXR image. This vector encapsulates the critical features identified by the CheXNet model, providing a concise yet informative representation of the image [2]. These feature vectors are subsequently used as inputs for the report generation model, ensuring that the generated reports are based on comprehensive and detailed image features. Through the application of CheXNet and these preprocessing techniques, we effectively harness advanced deep learning methods to extract meaningful features from CXR images, thereby facilitating the accurate and insightful generation of radiology reports.

4.4. Evaluation Metrics

In this research, the evaluation of the model's performance was conducted using BLEU (Bilingual Evaluation Understudy) scores. BLEU is a widely recognized metric for evaluating the quality of text generated by machine learning models, particularly in tasks such as machine translation and text generation. BLEU measures the correspondence between the machine-generated output and a reference output, with higher scores indicating better alignment between the two. We used BLEU-1, BLEU-2, BLEU-3, and BLEU-4 to capture n-gram overlaps between the predicted and actual sequences, thus providing a robust evaluation of the generated reports at different levels of granularity. BLEU-1 measures the unigram (single word) precision between the predicted and reference sequences. It evaluates the extent to which individual words from the predicted sequence appear in the reference sequence, without considering the order of the words. BLEU-2 extends this by considering bigram precision, which takes into account pairs of consecutive words, providing insight into how well the model captures short phrases or sequences of two words. BLEU-3 further evaluates the model's ability to generate contextually coherent sequences by considering trigrams (three consecutive words), while BLEU-4 assesses the precision of four-gram sequences, offering a more comprehensive evaluation of longer and more complex phrase structures. In this study, weighted BLEU scores were used to balance the contribution of different n-gram levels. For example, BLEU-2 applies equal weight to unigrams and bigrams, ensuring that both word-level and phrase-

level precision are accounted for. Similarly, BLEU-3 and BLEU-4 distribute the weights across unigrams, bigrams, trigrams, and four-grams, allowing for a more nuanced evaluation of the model's ability to generate coherent and contextually appropriate sequences. The evaluation process was conducted separately for both the cross-validation (CV) and test datasets. For each image-report pair, both the reference report and the model-generated report were preprocessed by removing punctuation and tokenizing the sequences. The BLEU scores were then calculated by comparing the n-grams between the predicted and reference sequences, allowing for an objective assessment of the model's performance. To ensure consistency, the average BLEU scores across all test samples were computed, providing a detailed view of the model's capability to generate accurate and coherent reports. To summarize the model's performance, an average BLEU score was calculated by taking the arithmetic mean of BLEU-1, BLEU-2, BLEU-3, and BLEU-4. This average score serves as a composite indicator of the model's performance across all n-gram levels, balancing both word-level precision and longer contextual coherence. This approach provides a holistic measure of the model's ability to generate high-quality reports, combining both accuracy at the word level and fluency across longer sequences.

Finally, to account for the varying lengths of the generated sequences, the BLEU scores were normalized by dividing the cumulative score by the sequence length. This normalization ensured that longer, more informative sequences were not penalized, preventing bias toward shorter sequences and providing a fair and balanced evaluation across all reports, regardless of their length. This normalization technique contributes to a more accurate representation of the model's true performance in generating coherent and contextually relevant outputs.

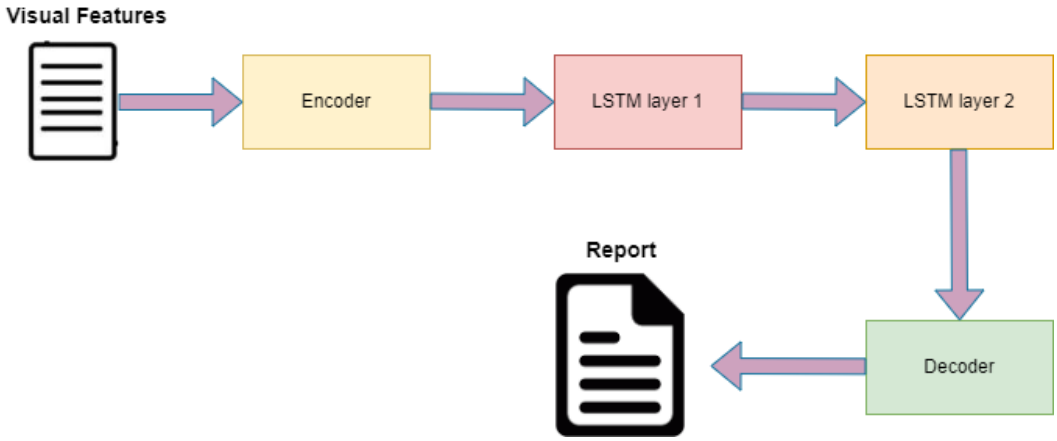

Figure 3: Medical Reports Automatic Generation Task.

4.5. Implementation and Parameter Settings

In this study, we employed an encoder-decoder Transformer architecture, combined with an LSTM layer, to automatically generate radiology reports from CXR images. Our proposed model is shown in Figure 3, the first stage of the model, the encoder, takes as input the image features extracted by CheXNet, a DenseNet-based CNN pre-trained weights on the CXR dataset. CheXNet transforms the input X-ray images into 1024-dimensional feature vectors. These vectors contain rich information regarding the visual characteristics of the images, such as anomalies, disease patterns, and other medically significant features. The CheXNet model was chosen due to its proven accuracy in identifying thoracic diseases, making it an ideal feature extractor for our report generation task. The extracted features are then passed into the Transformer encoder. The Transformer encoder consists of multiple layers of multi-head self-attention mechanisms and feed-forward networks. The self-attention mechanism allows the model to focus on different regions of the image feature set simultaneously, helping it capture global dependencies within the X-ray. Multi-head attention improves the model's ability to learn complex relationships by focusing on different parts of the input in parallel. To preserve the positional information of the image features, which the

Transformer lacks due to its architecture, positional encodings are added. These encodings ensure that the spatial structure of the image is maintained throughout the encoding process. In order to better capture sequential dependencies from the image features, we introduce an LSTM layer after the Transformer encoder. The LSTM is a type of recurrent neural network (RNN) known for its ability to handle long-term dependencies and sequential data effectively. In our model, the LSTM refines the encoded image features by considering temporal patterns across the feature sequence, which is critical for generating coherent and contextually relevant reports. The LSTM's role is particularly important as medical reports are typically sequential in nature, with certain observations depending on others.

After the LSTM has processed the image features, they are passed into the Transformer decoder, which is responsible for generating the radiology report. The decoder generates the report word by word in an autoregressive fashion, where each word is conditioned on both the previously generated words and the encoded image features. The decoder utilizes **a** masked multi-head self-attention mechanism to ensure that the generation process is sequential, preventing future words from influencing the current prediction. Additionally, the decoder uses a cross-attention mechanism to focus on the encoded image features while generating each word, ensuring that the generated text remains aligned with the content of the X-ray image. The model is trained using teacher forcing, where the ground truth words are provided during training to help the model learn the correct sequence of words. We optimize the model using the Adam optimizer with a learning rate of 0.001. The loss function used is categorical cross-entropy, which compares the predicted word probabilities with the ground truth words to guide the learning process. During inference, the model generates the report using beam search. Beam search allows the model to explore multiple possible word sequences at each generation step, selecting the sequence with the highest overall probability. This method improves the fluency and accuracy of the generated reports. Beam search was implemented with a beam width of 2, 5, and 7 in different experiments to assess its impact on the quality of the generated reports and to balance the trade-off between computational efficiency and generation quality.

In summary, our approach integrates the Transformer's powerful attention mechanisms with the sequential modelling capabilities of LSTM. The combination of these techniques enables the model to effectively process the complex visual data from CXRs and generate accurate, coherent radiology reports. This architecture ensures that the generated reports reflect both global and local features of the images, resulting in a system that is both robust and highly accurate for medical report generation. All experiments were conducted on an NVIDIA GPU model Tesla V100 with memory size 16GB to accelerate model training and reduce computation time.

Table 1

Comparison Study of Proposed Method with Previous Methods

Authors	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Li et al. [7]	0.438	0.298	0.208	0.151
Jing et al. [5]	0.455	0.288	0.205	0.154
Elaanba et al. [13]	0.27 (F1+F2)	-	-	-
Elaanba et al. [13]	0.26 (Frontal)	-	-	-
Elaanba et al. [13]	0.23 (Lateral)	-	-	-
Chen et al. [11]	0.470	0.304	0.219	0.165
Amjoud et al. [12]	0.479	0.359	0.219	0.160
our study	0.4636	0.4504	0.3754	0.3575

5. Results

In the results section, we present the performance of the model based on the BLEU evaluation metrics across multiple n-gram levels. The BLEU scores provide a detailed analysis of the model's ability to generate coherent and contextually relevant reports by comparing the predicted outputs with the reference sequences. For BLEU-1, which measures unigram precision and focuses on individual word matches, the model achieved a score of 0.4636. This indicates a relatively high level of word-level accuracy in the generated reports, reflecting the model's ability to correctly predict relevant terms that appear in the reference reports. The BLEU-2 score, which accounts for bigram precision and captures short phrases, was slightly lower at 0.4504. This suggests that while the model can accurately predict individual words, there is a slight reduction in performance when considering word pairs, indicating some challenges in maintaining short-term contextual coherence. As the evaluation extended to longer n-grams, the scores declined further. The BLEU-3 score, which assesses trigram precision, was 0.3754, indicating a more pronounced difficulty in generating contextually accurate sequences of three words. Similarly, the BLEU-4 score, which measures four-gram precision and reflects the model's ability to capture longer and more complex phrase structures, was 0.3575. The decreasing trend in BLEU scores from unigram to four-gram precision highlights the increasing complexity the model faces in generating longer, contextually accurate sequences. To summarize the overall performance, the average BLEU score, calculated as the mean of BLEU-1, BLEU-2, BLEU-3, and BLEU-4, was 0.4117. This composite score reflects the model's general performance across different n-gram levels, balancing both word-level precision and longer sequence coherence. While the model demonstrates strong performance in word-level accuracy, as indicated by the BLEU-1 score, its ability to generate longer, more coherent phrases and sentences is more challenging, as shown by the progressively lower BLEU scores for longer n-grams. These results suggest that while the model is effective at generating relevant words and short phrases, there is room for improvement in generating longer, contextually coherent sequences, which are critical for producing high-quality, fluent reports.

6. Discussion

The results of our study demonstrate the effectiveness of combining Transformer and LSTM architectures for generating radiology reports from CXR images, leveraging pretrained CheXNet for feature extraction. As indicated in Table 1, the model achieved a BLEU-1 score of 0.4636, highlighting its ability to accurately predict individual words relevant to medical reports. This performance suggests that the model is particularly adept at capturing important terms, which is critical for conveying key findings in radiology reports. However, the decline in performance across BLEU-2 to BLEU-4 metrics reflects the challenges the model faces in maintaining coherence in longer phrases and sentences. Specifically, the BLEU-4 score of 0.3575 points to difficulties in accurately generating complex, multi-word sequences. This limitation is expected, given the sequential nature of radiology reports, where specific findings and observations need to be described in detail and with context. The integration of the LSTM layer was designed to address such issues by capturing temporal dependencies, yet the model still struggles to consistently generate longer coherent sequences, indicating a potential area for improvement. The results also show that the inclusion of beam search during inference, with different beam widths, plays a significant role in balancing computational efficiency with the fluency and accuracy of the generated reports. By exploring multiple word sequences at each generation step, the model improves its output quality. Nevertheless, the progressively lower scores in higher n-grams suggest that there is room to further optimize this aspect of the model, possibly by exploring alternative decoding strategies or enhancing the sequential modeling of medical terminology.

A notable comparison can be made with the results of Elaanba et al. [7], who examined the impact of using frontal and lateral CXR images separately versus combining features from both views. In their study, the model achieved lower performance when using lateral (BLEU score 0.23) or frontal

(BLEU score 0.26) views alone, whereas a slight improvement was observed when combining both views (F1 + F2 score 0.27). This result underscores the value of multi-view image integration, as combining different perspectives of the same anatomical region provides richer feature representations, which is crucial for generating more comprehensive medical reports. Our approach similarly incorporates multiple image perspectives, which contributes to the overall model performance. However, the consistently higher BLEU scores in our study BLEU-4 of 0.3575 compared to Elaanba et al. [13] results suggest that integrating a Transformer-based attention mechanism with an LSTM layer may offer a more robust method for capturing both local and global features of the CXR images, compared to simpler architectures. Overall, the study demonstrates that while our model can effectively generate medical terms and short phrases, the challenge of producing fully coherent and contextually rich radiology reports remains. The comparison with Elaanba et al.'s work reinforces the importance of multi-view image integration, and future improvements could focus on refining how these views are processed. Additionally, more advanced attention mechanisms or domain-specific enhancements could be explored to further improve the generation of longer, contextually coherent sequences, ultimately enhancing the quality of the radiology reports.

7. Conclusions

This study presents a novel approach for generating radiology reports from chest X-ray (CXR) images by integrating a Transformer encoder with an LSTM layer. Leveraging CheXNet for feature extraction, our model effectively captures both global and local image features, while the LSTM enhances sequential modeling, which is crucial for producing coherent medical reports. The use of beam search during inference further improves the quality and fluency of the generated reports. Experimental results demonstrate that our model achieves competitive BLEU scores, with BLEU-1 at 0.4636 and BLEU-2 at 0.4504, indicating strong performance in capturing relevant medical terminology and generating coherent short phrases. Compared to previous methods, our model achieves higher BLEU-1 and BLEU-2 scores than the 0.438 reported by Li et al. and the 0.455 reported by Jing et al. Although the BLEU-3 (0.3754) and BLEU-4 (0.3575) scores declined, reflecting the challenge of generating longer, contextually rich sequences, they remain higher than those reported in related studies, such as Chen et al.'s model with a BLEU-4 score of 0.165. These results highlight our model's effectiveness at generating short, coherent sentences but also suggest areas for improvement in handling more complex sequences. While our approach shows significant promise, certain limitations should be noted. First, the model's reliance on a single-view CXR dataset may constrain its performance for complex cases that would benefit from multi-view imaging. Additionally, the model's computational complexity results in a relatively long runtime, which could impact real-time application feasibility. These limitations suggest opportunities for further enhancement. For future work, enhancing the LSTM layer with bidirectional LSTMs or exploring more efficient decoding strategies, such as top-k or nucleus sampling, could improve coherence in longer text sequences. Moreover, incorporating multi-view image analysis and domain-specific knowledge, such as clinical embeddings or expert annotations, may further refine report accuracy and contextual relevance. Overall, this study underscores that combining attention mechanisms with sequential modelling is a promising direction for advancing automated radiology report generation.

8. Declaration on Generative AI

During the preparation of this work, we used ChatGPT to assist with paraphrasing and improving sentence clarity, and Grammarly to assist with grammar and spelling checks. All AI-generated suggestions were critically reviewed and edited by the authors to ensure accuracy, originality, and

alignment with the publication's standards. The authors take full responsibility for the content and conclusions of this work.

9. References

[1] Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., & Summers, R. M. (2017). ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2097–2106). IEEE. doi:10.1109/CVPR.2017.369.

[2] Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., Lungren, M. P., & Ng, A. Y. (2017). CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning. arXiv preprint arXiv:1711.05225. Retrieved from https://arxiv.org/abs/1711.05225

[3] Johnson, A. E. W., Pollard, T. J., Greenbaum, N. R., Lungren, M. P., Deng, C.-y., Peng, Y., Lu, Z., Mark, R. G., Berkowitz, S. J., & Horng, S. (2019). MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*. Retrieved from https://arxiv.org/abs/1901.07042

[4] Alfarghaly, O., Khaled, R., Elkorany, A., Helal, M., & Fahmy, A. (2021). Automated radiology report generation using conditioned transformers. *Informatics in Medicine Unlocked*, 24, 100607. doi:10.1016/j.imu.2021.100607

[5] Jing, B., Xie, P., & Xing, E. P. (2018). On the automatic generation of medical imaging reports. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* (pp. 2577–2586). ACL. doi:10.18653/v1/P18-1240

[6] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* (Vol. 30, pp. 5998–6008). NeurIPS. Retrieved from https://papers.nips.cc/paper/7181-attention-is-all-you-need

[7] Li, C. Y., Liang, X., Hu, Z., & Xing, E. P. (2018). Hybrid retrieval-generation reinforced agent for medical image report generation. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI)* (pp. 2546–2552). IJCAI. doi:10.24963/ijcai.2018/354

[8] Liu, G., Hsu, T.-M. H., McDermott, M., Boag, W., Weng, W.-H., Szolovits, P., & Ghassemi, M. (2019). Clinically accurate chest X-ray report generation. In *Proceedings of the Machine Learning for Healthcare Conference* (pp. 249–269). PMLR. Retrieved from https://proceedings.mlr.press/v106/liu19a.html
.

[9] Zhang, Y., Wang, X., Xu, Z., Yu, Q., Yuille, A., & Xu, D. (2020). When radiology report generation meets knowledge graph. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 34, No. 1, pp. 12910–12917). AAAI. doi:10.1609/aaai.v34i01.7006

[10] Lovelace, J., & Mortazavi, B. (2020). Learning to generate clinically coherent chest X-ray reports. *IEEE Journal of Biomedical and Health Informatics*, 24(11), 3382–3392. doi:10.1109/JBHI.2020.2997038

[11] Chen, Z., Song, Y., Chang, T.-H., & Wan, X. (2020). Generating radiology reports via memory-driven transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1439–1449). ACL. doi:10.18653/v1/2020.emnlp-main.113

[12] Amjoud, A. B., & Amrouch, M. (2021). Automatic generation of chest X-ray reports using a transformer-based deep learning model. In *Proceedings of the 2021 Fifth International Conference on Intelligent Computing in Data Sciences (ICDS)* (pp. 43–48). IEEE. doi:10.1109/ICDS54393.2021.9664861

[13] Elaanba, A., Ridouani, M., & Hassouni, L. (2024). Transformer-based model for radiology text reports generation from frontal and lateral chest X-ray images. *International Journal of Computer Information Systems and Industrial Management Applications*, 16, 345–356.

[14] Mohamed, G., Eldib, H., & Sharkas, M. (2021). Seizure prediction using two-dimensional discrete wavelet transform and convolution neural networks. In *Proceedings of the International Workshop on Informatics & Data-Driven Medicine* (pp. 78–83). IEEE. doi:10.1109/IWIDDM.2021.9456135

[15] Huang, G., Liu, Z., van der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 4700–4708). IEEE. doi:10.1109/CVPR.2017.243

[16] Iandola, F. N., Han, S., Moskewicz, M. W., Ashraf, K., Dally, W. J., & Keutzer, K. (2016). SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5 MB model size. *arXiv preprint arXiv:1602.07360*. Retrieved from https://arxiv.org/abs/1602.07360

[17] Baltruschat, I. M., Nickisch, H., Grass, M., Knopp, T., & Saalbach, A. (2019). Comparison of deep learning approaches for multi-label chest X-ray classification. *Scientific Reports*, 9(1), 6381. doi:10.1038/s41598-019-42294-8

[18] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R., & Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)* (pp. 2048–2057). PMLR. doi:10.48550/arXiv.1502.03044

[19] Cornia, M., Stefanini, M., & Baraldi, L. (2020). Meshed-memory transformer for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 10578–10587). IEEE. doi:10.1109/CVPR42600.2020.01059

[20] Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (pp. 311–318). ACL. doi:10.3115/1073083.1073135

[21] Wiseman, S., Shieber, S. M., & Rush, A. M. (2017). Challenges in data-to-document generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 2253–2263). ACL. doi:10.18653/v1/D17-1239

[22] Demner-Fushman, D., Kohli, M. D., Rosenman, M. B., Shooshan, S. E., Rodriguez, L., Antani, S., Thoma, G. R., & McDonald, C. J. (2016). Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2), 304–310. doi:10.1093/jamia/ocv080

[23] Soni, R. K. (2021). Indiana University — Chest X-Rays automated report generation. [Online]. Available: https://rohansoni-jssaten2019.medium.com/indiana-university-chest-x-rays-automated-report-generation-38f928e6bfc2

[24] Clark, A. (2008). ElementTree: An API for XML parsing and generation. *Python Library Documentation*. [Online]. Available: https://docs.python.org/3/library/xml.etree.elementtree.html

[25] Bradski, G. (2000). The OpenCV Library. *Dr. Dobb's Journal of Software Tools*. [Online]. Available: https://opencv.org/about/

[26] Barzilai, E., & C. G. (2020). Handling missing data in textual datasets using contextual embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5434–5439). ACL. doi:10.18653/v1/2020.acl-main.485

[27] Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly Media.