

A multi-modal feature fusion-based approach for Chest X-ray Report Generation

Fatima CHEDDI

Smart Systems Laboratory
ENSIAS, Mohammed V University in Rabat,
Rabat, Morocco
Fatima.chedi@gmail.com

Ahmed HABBANI

Smart Systems Laboratory
ENSIAS, Mohammed V University in Rabat,
Rabat, Morocco
IBISC Laboratory
ParisSaclay University
Evry University, France
CNRS
Bordeaux INP
LaBRI, Univ. Bordeaux, Talence, France
ahmed.habbani@ensias.um5.ac.ma

Hammadi Nait-Charif

National Center for Computer Animation
Bournemouth University
United Kingdom
hncharif@bournemouth.ac.uk

Abstract— Given the growing dependence on medical imaging, there is a significant requirement for automated report generation, which can save the radiologist's time and reduce the possibility of diagnostic errors. Existing approaches face various difficulties, including insufficient professionalism, a variety of diseases, and fluency in reports. These problems are the result of the use of an encoder-decoder deep learning architecture to establish a uni-directional image-to-report relationship and neglect the bidirectional connections between images and reports, making it challenging to establish the intrinsic medical correlations between them. To this end, we propose a novel approach for chest radiology report generation based on multi-modal feature fusion. Our method uses textual and visual features that are taken from medical chest X-ray images and their real reports. Firstly, we use a vision transformer to extract visual features from medical images; on the other hand, we use the Word2Vec model to extract semantic features from textual medical reports. Additionally, we employ advanced techniques such as channel attention networks and cross-modal information fusion modules to enhance the quality and coherence of the generated reports. We have evaluated our proposed approach on two publicly available chest X-ray datasets, IU X-ray and NIH. The results show that our approach outperforms state-of-the-art methods. Particularly in the ROUGE metric and BLEU metric.

Keywords—Chest X-ray report generation, Multi-modal learning, medical image, Deep learning, Transformers, Medical report

I. INTRODUCTION

In recent years, with the development of artificial intelligence in medical practices, automatic report generation based on deep learning has become a popular research topic.

Chest X-ray (CXR) radiographic imaging plays a pivotal role in the early diagnosis of various thoracic diseases [1]. Radiologists carefully examine each section of a chest X-ray image, documenting in written reports any findings and results. These reports summarize the diagnostic analysis and conclusions derived from the radiological image and act as a standard method of communication between radiologists and doctors [2]. Furthermore, manual interpretation is time-consuming and requires a significant portion of a specialist's time and effort [3]. This emphasizes the need for automated methods to assist radiologists in their work. Deep learning-based automated diagnostic report generation can reduce the time required to write reports and minimize the possibility of errors in diagnosis [4], thus improving work efficiency. In recent years, advanced searches in deep learning and natural language processing (NLP) have paved the way for automated CXR report generation. Current methods typically use an encoder-decoder model [5] [6], where the encoder utilizes convolutional neural networks (CNN) to extract visual features from the radiology image, while the decoder employs recurrent neural networks (RNN) to produce text descriptions for the reports.

To improve the models mentioned above, the researchers applied a transformer [7] [8] to the task of medical report generation. In comparison to current report generation approaches, the transformers model exhibits enhanced feature learning capabilities, as presented in these works [9] and [10], as they can weigh the relationships between each node in the input data and are adept at capturing long-range semantic associations. Also, they present higher interaction with multimodal data due to the multi-head attention module. Furthermore, transformers use self-attention mechanisms to improve parallelization and comprehension of word links in a series of words. Transformers provide faster and more efficient

learning because they do not rely on recurrence and add more context to the network compared to RNNs.

The existing research shows that the methods based on the improved transformer models and global visual feature have achieved excellent performance in the task of automatic report generation, but at the same time, these methods still have some limitations when it comes to identifying and characterizing rare and complex diseases due to a lack of in-depth understanding of medical terminology, anatomical structures, or lesion characteristics. Furthermore, the model's description of diseases lacks professionalism due to its limited knowledge of medical terms, anatomical structures, and lesion characteristics, producing text that is difficult to understand or not fluent, which negatively impacts the report's usability.

To address these limitations, we propose a novel method for CXR report generation based on multi-modal feature fusion techniques, termed MM-CXRG. Our approach integrates both visual and textual modalities using deep learning algorithms, including convolutional neural networks (CNNs) and improved transformers, aiming to improve the quality of automated medical report generation. Our approach focuses on solving the problems of disease recognition accuracy, disease description professionalism, and report text fluency, thereby decreasing the burden for clinicians in the face of report writing and assisting them in achieving accurate diagnoses.

II. RELATED WORK

Over several years, Convolutional Neural Networks (CNNs) have played a pivotal role in medical image analysis, relying on a straightforward and implementable architecture [11] and [12]. In contrast, while these studies [5] and [13] used the CNN-RNN framework, more recent works have adopted the power performance of the Transformer [8], which employs the mechanism named self-attention [34].

On the other hand, several searches have explored reinforcement learning and contrast learning techniques to enhance models' report generation capability. For example, [13] presented a framework that compares input images with normal ones to extract contrast information, improving the characterization of abnormal regions. In [14], Wang et al. proposed a self-reinforcement method where image-report pairs are fed into an image-text matching branch, providing feedback on alignment to enhance report accuracy. Kaur [15] proposed a method integrating attention and reinforcement learning strategies to improve the relevance of the generated report text to medical images. Also, Li et al. proposed an HRGR model [16] that uses reinforcement learning to create radiology reports by hand-assembling a manually extracted template database. However, these approaches do have some challenges, including the need for large amounts of labeled data to train efficiently. Secondly, the complexity of RL models increases the time required for training. Also, the possibility of overfitting and

integrating RL with attention mechanisms or contrast learning adds further complexity and limits the power of the system to interpret complex cases.

In order to increase the quality of the radiology reports that are generated, other approaches are proposed that incorporate additional past knowledge into the model. For example, Shuxin Yang introduced a method named M2KT [17], which combines two separate modules. Learned knowledge base and multi-modal alignment, another approach introduced by Li et al., named Knowledge-driven Encode, Retrieve Paraphrase (KERP) [18], which splits medical reports into two parts: the learning of explicit medical abnormality graphs and following natural language modeling. This method uses three modules: Encoder, Retrieve, and Paraphrase, which edit the templates based on particular cases. Also, another method is proposed by Liu et al. named the Posterior-and-Prior Knowledge Exploring-and-Distilling approach (PPKED)[19]. The model utilizes embeddings representing common abnormalities observed in training set reports, encoded reports with similar visual features to the current CXR, and knowledge graph embeddings reflecting prevalent abnormalities in the training data. In the same field, another approach named R2Gen, proposed by Chen et al. [20], incorporates a relational memory-driven mechanism within the decoder to grasp the sequencing of words or sentences. followed by the development of the 'Cross-modal Memory Network' (CMN). Recently, a novel mechanism has been presented using the multi-modal fusion data utilized in these works [21], [22], and [23]. Its objective is to improve the interaction between different modalities of data to improve the quality and accuracy of the generated reports and to reduce errors compared to strategies that rely on one type of data. By incorporating multiple modalities, the model can capture a more comprehensive understanding of the medical image, including anatomical structures, lesion characteristics, and associated clinical findings.

However, these works also come with some limitations. The models might be limited to specific fields, and they may need significant modifications for new applications. In addition, the quality of the training data is crucial for ensuring the accuracy of reports. Any inaccuracies or biases in a data source, whether it be images or medical records, can result in reports that are not reliable across all references.

III. METHODOLOGY

A. Architecture overview

The proposed method for image chest X-ray report generation, as illustrated in Fig. 1, is divided into three modules: visual and text feature extraction, multimodal feature fusion, and decoder report generation. The visual and text feature extraction module turns two types of data (images and texts) into semantic information. For feature image extraction, we resized each image to 224x224 pixels and then used a CNN as the backbone network with a vision transformer. On the other hand, we use Word2Vec to extract the textual feature

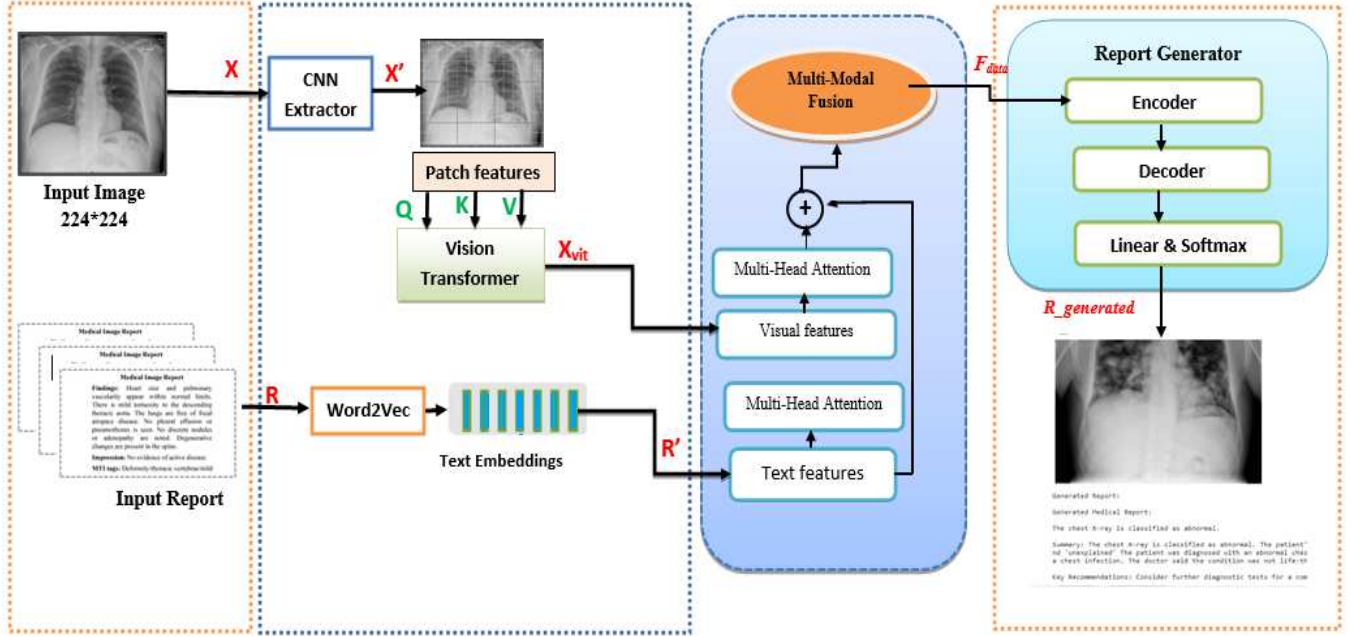


Fig. 1. Overall framework

information. We combine these features and store them in a vector with multiple dimensions, forming a distributed representation, which we then feed into the feature fusion network. This fused feature vector combines both detailed local features and global contextual information. This module uses attention mechanisms to structure a unified feature representation matrix. The final step utilizes the output to produce radiological reports through the Decoder BART, a bidirectional and transformer language-based system.

B. Image Feature Extraction

For this task, we employ a hybrid model integrating Convolutional Neural Networks and Vision transformers. The process starts with inputting the chest X-ray image. The image is passed through a set of convolutional layers, where each layer employs multiple convolutional filters. The output is a set of feature maps that represent the image's high-level local features. The CNN output is divided into non-overlapping patches. The chest x-ray image with a size 224x224 might be divided into 16x16 patches, which produces 196 patches. Each patch is transformed into a vector and then projected into a high-dimensional embedding space using a linear projection to obtain a sequence of patch embeddings. Then we fed this sequence into the ViT model to extract contextual information and global dependencies. This approach combines the strengths of CNNs for local feature extraction with transformers, which incorporate a self-attention mechanism for global context modeling. This approach improves the handling of both local and global information, thereby optimizing performance. The formula outlined in Equations (1)– (4) explains this process.

In this work, we utilize the classical convolutional neural network (CNN) ResNet-50 [24] model as the image feature extractor due to its superior performance. The following metrics illustrate the process of extracting image modal data (Image1, Image2, ..., ImageN) using the ResNet-50 model.

- Feature Extraction using ResNet:

Let X_{ResNet} represent the output feature maps obtained from ResNet:

$$X' = ResNet(Image_x) \quad (1)$$

- Patch Embedding (for Vision Transformer): Linearize the feature maps into patches:

$$X_{patch} = Linearize(X') \quad (2)$$

- Positional Encoding (for Vision Transformer): Add positional encodings to the patch embeddings:

$$X_i = X_{patch} + PE_i \quad (3)$$

- Transformer Encoder (for Vision Transformer): Pass the combined patch embeddings through transformer encoder layers:

$$X_{vit} = ViT(X_i) \quad (4)$$

C. Textual Feature Extraction

The textual report data includes a description, comparison, finding, etc. of the patient's condition. As shown in Fig.2 and equation (5), To extract the semantic information from each text in the report, we use the Word2Vec Model [35], a popular word embedding technique for extracting meaningful features from

medical reports. Word2Vec facilitates the transformation of text data into dense vector representations, capturing semantic relationships between words and enabling more effective analysis.

$$R' = \text{Word2vec}(R) \quad (5)$$

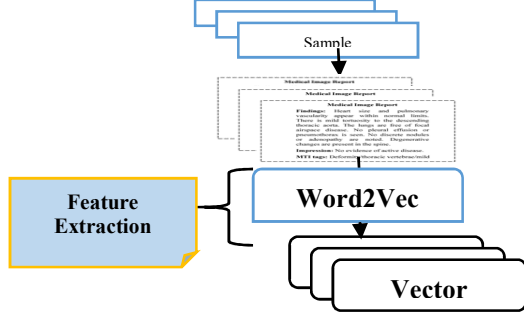


Fig. 2. Text Feature Extraction

D. Multimodal Fusion Module (MFM)

After extracting the features from both the previous module's visual features and text representations, we projected them to a multi-modal fusion attention module in order to concatenate and acquire a multi-modal representation as shown in equation (6). For mapping and separating the resulting features into latent data representations, we adopt the technique of multi-layer perception.

$$Fdata = \text{Fusion} (XViT, R') \quad (6)$$

E. Report Generator (RG)

For this task, as presented in equation (7) the obtained multi-modal features serve as the query and key inputs for the encoder-decoder model to generate the final chest X-ray report. The encoder-decoder model adopted in this RG is named Bidirectional and Auto-Regressive Transformer BART [25]. This model is a combination of a BERT [26] based bidirectional encoder and a GPT [27] model based on self-attention and an autoregressive decoder.

$$R_{\text{generated}} = \text{BART}(F_{\text{data}}) \quad (7)$$

IV. EXPERIMENTAL SETUP

Our proposed approach for automated report generation was evaluated and validated on the IU [28] and NIH [36] datasets. In this subsection, we prese a detailed description of these public datasets and use evaluation metrics.

A. Datasets

1) Indiana University (IU)

One of the most popular datasets is the Indiana University (IU) chest X-ray collection [28], which includes 7470 multiple

CXR images and 3995 reports. This publicly available dataset was developed to train and test algorithms for tasks including categorization, summary production, and report generation using medical X-ray images. Each report includes impressions, reports, tags, and patient information.

2) NIH Chest X-ray :

This dataset [36], released by the National Institutes of Health (NIH), is a comprehensive collection of over 100,000 frontal-view chest X-ray images obtained from more than 30,000 persons. Each image is annotated with up to 14 thoracic disease categories, such as pneumonia, fibrosis, and emphysema.

B. Evaluation metrics:

To assess the quality of our approach, we utilize various automatic evaluation metrics including BLEU [29], METEOR [30], ROUGE L [31] . in order to compare the generated report with the truth reports.

Initially, BLEU-1 to BLEU-4 scores were computed to compute n-gram precision. The higher value of BLEU represents a closer word-level similarity between the generated and truth reports. Additionally, the ROUGE-L score was employed to assess the longest common subsequence, providing insights into the recall and precision of the generated text. Finally, the METEOR metric indicates the similarity between two sentences based on synonyms, and repetitions, which can accurately reflect the quality of generated text.

V. Results and discussion

In this section, we provide an in-depth analysis of the results obtained from our implementation including quantitative and qualitative analyses. Lastly, a comparison study is applied between our proposed model and some of the existing approaches.

A. Quantitative Results

Table I: Quantitative results of our method

Dataset	B-1	B-2	B-3	R- L	METEOR
IU	0.482	0.312	0.229	0.381	0.198
NIH	0.401	0.332	0.221	0.409	0.393

Our proposed method uses a transformer model, trained on medical images. We used the IU and NIH datasets to evaluate our model using automatic evaluation metrics such as Bleu-n,

ROUGE_L, and METEOR. Table I illustrates the experimental results.

During the training phase, the model's performance indicators gradually improve and eventually converge. As shown in Table I, the highest scores for BLEU-1, BLEU-2, BLEU-3, BLEU-4, ROUGE-L, and METEOR are 0.482, 0.312, 0.229, 0.381, and 0.198. To validate the effectiveness and generalization of our model, we compared it with existing methods based on transformers and CNN.

B. Qualitative Results

Medical X-ray image	Ground truth report	Generated report Using BART
	Findings: The chest X-ray reveals patchy infiltrates in the right lower lobe with air bronchograms, indicative of pneumonia. There is also evidence of consolidation in the same area. No pleural effusion or pneumothorax is observed. there is mild tortuosity to the descending thoracic aorta. The heart size appears normal. Impression: Right lower lobe pneumonia with patchy infiltrates and consolidation. No evidence of pleural effusion or pneumothorax. Normal heart size.	Findings: The chest X-ray shows patchy infiltrates in the right lower lobe, suggesting the presence of pneumonia. Air bronchograms are visible in the affected area. There is no evidence of pleural effusion or pneumothorax. The heart size appears normal. Impression: Right lower lobe pneumonia with patchy infiltrates. Absence of pleural effusion or pneumothorax. Normal heart size.

Fig. 3. Visualisation of medical report generation

Fig. 3 displays the input medical image, real reports, and the reports generated by our model. Observing the experimental results, it is evident that the multi-modal model can focus on abnormal areas. For example, the model correctly identifies diseases such as "presence of pneumonia" and can generate descriptions like "air bronchograms" for disease status, demonstrating its effectiveness. However, it struggles to accurately identify early or less apparent conditions, such as "there is mild tortuosity to the descending thoracic aorta," and there are still some differences in the way the generated reports describe diseases compared to real reports. Our approach can accurately generate specialized medical terms such as "air bronchograms" and "pneumonia," improving the professionalism and logical coherence of the reports. It also provides a more comprehensive description of normal organs. It also demonstrates the effectiveness of the multi-modal fusion strategy. Also, the reports generated by our model can

accurately describe the position of "normal heart size" but do not explain that "there is mild tortuosity to the descending thoracic aorta." Additionally, the reports generated are more consistent in length with real reports, and the sentences are smoother and more fluent. Thus, the accuracy of the model-generated reports has improved, clearly expressing the specific location and attributes of most lesions.

C. Comparison and discussion:

Table II: Performance Comparison results.

Data-Set	Method	Bl 1	B 2	B 3	Rouge L	MET-EOR
IU Xray	R2Gen	0.470	0.304	0.219	0.371	0.187
	HRGR	0.438	0.298	0.208	0.322	----
	CMN	0.475	0.309	0.222	0.375	0.165
	PPKED	0.483	0.315	0.224	0.376	0.191
	ITM	0.496	0.319	0.241	0.377	
	Ours	0.482	0.312	0.251	0.381	0.198
NIH	X-Swin-Gen	0.396	0.329	----	0.404	0.393
	ours	0.409	0.335	0.231	0.410	0.396

To illustrate the efficacy of our approach, we conducted comparative experiments with existing models using the datasets mentioned below. Specifically, we choose models such as HRGR [16], R2GEN [20], CMN [32], PPKED [19], ITM [33], and XraySwinGen [37] as comparison models. The results of these comparative experiments are outlined in Table II. From what we know, it's clear that our method does better on the IU X-ray dataset than the R2Gen, HRGR, and CMN models when using the overall assessment metrics (BLEU-3, METEOR, and CIDER). Specifically, the ROUGE-L metric achieved a 0.381, whereas the METEOR metric demonstrated a 0.198% improvement. However, the PPKED and ITM both attained the highest mean BLEU-1 and BLEU2 scores. This indicates the efficacy of utilizing a transformer decoder in the generation of medical reports. On the NIH dataset, our method achieves a BLEU-1 score of 0.409 and a BLEU-2 score of 0.335, which are slightly better than the XraySwinGen method. Additionally, our method achieves the highest ROUGE-L score of 0.410, demonstrating its effectiveness in generating coherent and contextually relevant reports. Our METEOR score of 0.393 matches that of XraySwinGen, further validating the robustness of our approach across different datasets.

Both the qualitative and quantitative results show that the proposed approach notably enhances the model's ability to

detect chest diseases and accurately analyze their appearance using the hybrid strategies, resulting in a more accurate lesion description .

This superior performance can be attributed to the model's capability to use strategies like the multi-head attention mechanism in transformers, the multimodal fusion module based on attention mechanisms, the Word2Vec extractor, and the BART-based generator. This allows the model to understand and generate synonyms and terms used by different radiologists. Also, training on a large dataset such as IU chest x-rays and NIH, which includes reports from multiple radiologists with different writing styles, helps the model prioritize critical information over stylistic variations. The ability to handle variations in writing styles and vocabulary improves the robustness and reliability of the model. It ensures that our model can address situations when radiologists have different writing styles and vocabulary while describing the same radiological image. It also maintains high accuracy and quality in the output reports.

Additionally, while this approach performs well, it still needs improvement. For future work, we will begin by integrating self-refining strategy and data augmentation. Secondly, we will employ other datasets in the field of brain MRI and ultrasound in the experimentation to enhance the training process.

VI. CONCLUSION AND PERSPECTIVES

Our study presents A new approach for Chest X-ray Report Generation by combining visual and textual data with advanced deep learning models, utilizing multi-modal feature fusion and the BART Generator applied on two datasets. Our methodology addresses key obstacles in the generation of medical reports, such as raising the precision of disease identification, improving the quality of disease detection, and assuring the coherence of the generated reports. Our method efficiently achieves a significant result by using multi modal fusion strategy and the BART as the decoder. The experimental results show that our approach surpasses existing methods in terms of several evaluation metrics, highlighting its potential to aid radiologists in producing precise and dependable diagnostic reports. Our future work will prioritize the improvement of multi-modal feature fusion techniques. Also, other datasets will be explored in future experimentation. This will improve the reliability and applicability of our methodology in other diseases.

REFERENCES

- [1] Raoof, S., Feigin, D., Sung, A., Raoof, S., Irugulpati, L., & Rosenow, E. C., "Interpretation of Plain Chest Roentgenogram. Chest", 141(2), 545-558, 2012.
- [2] Monshi, M. M. A., Poon, J., & Chung, V., "Deep learning in generating radiology reports: A survey. Artificial Intelligence in Medicine", 106, 101878, 2020.
- [3] Kaur, N., Mittal, A. & Singh, G. "Methods for automatic generation of radiological reports of chest radiographs: a comprehensive survey", *Multimed Tools Appl* 81, 13409–13439, 2022.
- [4] Y. Zhang, X. Wang, Z. Xu, Q. Yu, A. Yuille, and D. Xu, "When Radiology Report Generation Meets Knowledge Graph", *AAAI*, vol. 34, no. 07, pp. 12910-12917, Apr. 2020.
- [5] Baoyu Jing, Zeya Wang, and Eric Xing. , "Show, Describe and Conclude: On Exploiting the Structure Information of Chest X-ray Reports" In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6570–6580, Florence, Italy. Association for Computational Linguistics, 2019.
- [6] Baoyu Jing, Pengtao Xie, and Eric Xing. , "On the Automatic Generation of Medical Imaging Reports", In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2577–2586, Melbourne, Australia. Association for Computational Linguistics, 2018.
- [7] Von Luxburg, Ulrike, et al. "Advances in neural information processing systems 30." ,31st annual conference on neural information processing systems ,Long Beach, California, USA, 2017.
- [8] Vaswani, A., Shazeer, N.M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., & Polosukhin, I. , "Attention is All you Need", *Neural Information Processing Systems*, 2017.
- [9] K. Han et al., "A Survey on Vision Transformer," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 1, pp. 87-110, Jan. 2023.
- [10] He, X., Yang, Y., Shi, B., & Bai, X. , "VD-SAN: Visual-Densely Semantic Attention Network for Image Caption Generation", *Neurocomputing*, 328, 48-55, (2019).
- [11] Santosh Kumar, Sachin Kumar Gupta, Vinit Kumar, Manoj Kumar, Mithilesh Kumar Chaube, Nenavath Srinivas Naik, "Ensemble multimodal deep learning for early diagnosis and accurate classification of COVID-19, *Computers and Electrical Engineering*", Volume 103 , 108396, ISSN 0045-7906,2022.
- [12] Litjens, Geert, et al. "A survey on deep learning in medical image analysis." *Medical image analysis* 42: 60-88, 2017.
- [13] Y. Zhang, X. Wang, Z. Xu, Q. Yu, A. Yuille, and D. Xu, "When Radiology Report Generation Meets Knowledge Graph", *AAAI*, vol. 34, no. 07, pp. 12910-12917, Apr. 2020.
- [14] Wang, X., Zhang, Y., Guo, Z., & Li, J., "TMRGM: A Template-Based Multi-Attention Model for X-Ray Imaging Report Generation", *Journal of Artificial Intelligence for Medical Sciences*, 2021.
- [15] Navdeep Kaur, Ajay Mittal, "CADxReport: Chest x-ray report generation using co-attention mechanism and reinforcement learning, *Computers in Biology and Medicine*, Volume 145, 2022.
- [16] Christy Y. Li, Xiaodan Liang, Zhiting Hu, and Eric P. Xing. , "Hybrid retrieval-generation reinforced agent for medical image report generation", In *Proceedings of the 32nd International Conference on Neural Information Processing Systems* . Curran Associates Inc., Red Hook, NY, USA, 1537–1547, 2018.
- [17] Yang S, Wu X, Ge S, Zheng Z, Zhou SK, Xiao L. , "Radiology report generation with a learned knowledge base and multi-modal alignment", *Med Image Anal*. 2023 May;86:102798, Epub Mar 23, 2023.
- [18] Li, C. Y., Liang, X., Hu, Z., & Xing, E. P. , "Knowledge-Driven Encode, Retrieve, Paraphrase for Medical Image Report Generation", *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01), 6666-6673, 2019.
- [19] F. Liu, X. Wu, S. Ge, W. Fan and Y. Zou, "Exploring and Distilling Posterior and Prior Knowledge for Radiology Report Generation", *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA, pp. 13748-13757, 2021.
- [20] Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. , "Generating Radiology Reports via Memory-driven Transformer", In *Proceedings of the 2020 Conference on Empirical Methods in Natural*

Language Processing (EMNLP), pages 1439–1449, Online. Association for Computational Linguistics, 2020.

- [21] Yu Pan, Li-Jun Liu, Xiao-Bing Yang, Wei Peng, Qing-Song Huang, “Chest radiology report generation based on cross-modal multi-scale feature fusion”, *Journal of Radiation Research and Applied Sciences*, Volume 17, Issue 1, 2024.
- [22] Aksoy, N., Sharoff, S., Baser, S., Ravikumar, N., & Frangi, A. F., “Beyond images: An integrative multi-modal approach to chest x-ray report generation”, *Frontiers in Radiology*, 4, 1339612, 2024.
- [23] Yun Tan, Chunzhi Li, Jiaohua Qin, Youyuan Xue, Xuyu Xiang, and Costa Gianni. “Medical Image Description Based on Multimodal Auxiliary Signals and Transformer”, *Int. J. Intell. Syst.* 2024.
- [24] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition", *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, pp. 770-778, 2016.
- [25] Lewis M, Liu Y, Goyal N, Ghazvininejad M, Mohamed A, Levy O, Stoyanov V, Zettlemoyer L., "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension", In: *Proceedings of the 58th annual meeting of the association for computational linguistics*. 2020.
- [26] Devlin, J., Chang, M., Lee, K., & Toutanova, K., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". *North American Chapter of the Association for Computational Linguistics*, 2019.
- [27] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever., "Improving language understanding by generative pre-training, 2018.
- [28] Demner-Fushman D, Kohli MD, Rosenman MB, Shooshan SE, Rodriguez L, Antani S, Thoma GR, McDonald CJ., "Preparing a collection of radiology examinations for distribution and retrieval", *Journal of the American Medical Informatics Association : JAMIA*, 23(2), 304–310, 2016.
- [29] Papineni, Kishore, et al. "Bleu: a method for automatic evaluation of machine translation." *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002.
- [30] Lavie, A. "Agarwal. A.," METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments." *Proceedings of Workshop on Statistical Machine Translation at the 45th Annual Meeting of the Association of Computational Linguistics*, 2007.
- [31] Lin, Chin-Yew. "Rouge: A package for automatic evaluation of summaries." *Text summarization branches out*, 2004.
- [32] Chen, Z., Shen, Y., Song, Y., & Wan, X., "Cross-modal memory networks for radiology report generation", In *ACL-IJCNLP 2021 - 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing, Proceedings of the Conference* (pp. 5904–5914), 2021.
- [33] Wang, Z., Zhou, L., Wang, L., & Li, X., "A self-boosting framework for automated radiographic report generation", *IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 2433–2442, 2021.
- [34] Fenglin Liu, Changchang Yin, Xian Wu, Shen Ge, Ping Zhang, and Xu Sun., "Contrastive Attention for Automatic Chest X-ray Report Generation", In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 269–280, Online. Association for Computational Linguistics, 2021.
- [35] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.
- [36] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, R.M. Summers, "ChestX-ray8: hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases", *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, pp. 2097–2106, 2017.
- [37] Veras Magalhães G, L de S Santos R, H S Vogado L, Cardoso de Paiva A, de Alcântara Dos Santos Neto P., "XRaySwinGen: Automatic medical reporting for X-ray exams with multimodal model", *Heliyon*. Mar 12;10(7):e27516, 2024.