A multimodal framework for explainable chest X-ray report generation

Hamza Chehili[1,2], Nourhene Bougourzi[1], Raida Malak Makhlouf [1], Hadjer Taib[1], Mustapha Bensaada[1]
[1]University frères Mentouri, Constantine 1, Constantine 25000, Algeria
[2]LIRE Laboratory, Constantine 2 University, Constantine 25000, Algeria

Article Info

Article history:

Received Oct 14, 2024
Revised Jan 13, 2026
Accepted Feb 28, 2026

Keywords:

Chest X-ray report generation
Explainable AI
Grad-CAM
Low-rank adaptation
Multimodal learning

ABSTRACT

Chest X-ray (CXR) interpretation remains a challenging task due to overlapping anatomical structures, variability in disease presentation, and increasing clinical workload. Existing automated report-generation models provide promising results but often lack explicit interpretability, limited clinical alignment, and insufficient comparative evaluation with established baselines. This study proposes an explainable multimodal framework that combines a dual CNN encoder (ResNet-50 and EfficientNet-B0) with the Gemma-3 1B language model fine-tuned using low-rank adaptation (LoRA). Visual explanations are produced through Gradient-weighted Class Activation Mapping (Grad-CAM) to enhance transparency in the decision process. Unlike prior image-to-text pipelines, our approach follows a findings-guided paradigm and integrates both visual and textual cues during generation. Experiments conducted on public datasets demonstrate consistent improvements over representative vision-language baselines reported in recent literature, with notable gains in BLEU, ROUGE, METEOR, and BERTScore. Generated reports show improved factual completeness and clinically relevant region-level attention. Limitations include the absence of evaluation against emerging foundation models and the need for anatomical- level explainability metrics. Future work will extend benchmarking to models such as M2-Transformer, MedCLIP-GPT, and R2Gen, and will explore clinical validation in real-world workflows.

Corresponding Author:

Hamza Chehili
University frères Mentouri, Constantine 1
Constantine 25000, Algeria
Email: h.chehili@umc.edu.dz

1. INTRODUCTION

Chest X-rays (CXRs) are the most widely used radiological examination worldwide and play a central role in diagnosing thoracic diseases, including pneumonia, tuberculosis, heart failure, and malignancies [1]. Despite standardized acquisition protocols such as posteroanterior (PA) and lateral views, CXR interpretation remains difficult due to anatomical superposition and the inherent limitations of 2D projections, even when following structured mnemonics such as ABCDEFGHI [2]. The growing global demand for imaging, combined with the shortage of trained radiologists, has intensified diagnostic delays and increased the risk of interpretive errors associated with fatigue and high workloads [3]. Report variability, non-standardized free-text descriptions, and limited machine-readability further complicate clinical decision-making and downstream data extraction [4].

Artificial intelligence (AI) is increasingly investigated to mitigate these challenges, particularly in automated radiology report generation [5]. Unlike traditional classification, report generation requires models capable of connecting visual abnormalities with clinically coherent language, necessitating the integration of computer vision and natural language processing [6], [7]. Recent advances in multimodal large language models (LLMs) have significantly improved medical image understanding [8]. Convolutional neural networks (CNNs) remain foundational for feature extraction in radiology, with architectures such as ResNet and EfficientNet offering strong representational performance, especially when combined through ensemble learning and transfer learning strategies [9]. However, the clinical adoption of these systems remains limited due to a lack of interpretability and the risk of hallucinated text, which can undermine trust in automated reports [10], [11].

Recent multimodal frameworks, including XrayGPT, CXR-LLAVA, ELIXR, LiteGPT, and RoentGen, demonstrate substantial progress in vision-language modeling for CXRs (Table 1). These systems integrate visual encoders with LLMs to perform summarization, question answering, classification, or image synthesis. Yet, most follow a direct image-to-text paradigm, offer limited explainability, and do not incorporate a findings-guided workflow, which is standard in clinical reporting.

To address these limitations, this study proposes a novel explainable framework that integrates four complementary innovations. First, we introduce a findings-as-input paradigm that aligns more closely with real radiological workflows [12]. Second, we employ a dual CNN ensemble combining ResNet-50 and EfficientNet- B0 to improve visual representation quality [13]. Third, gradient-weighted class activation mapping (Grad- CAM) is used to visualize model attention patterns and enhance interpretability [14]. Finally, the Gemma-3 1B language model is efficiently fine-tuned using low-rank adaptation (LoRA), enabling domain adaptation while maintaining computational efficiency [15]. The contribution of this work lies in bridging performance and explain.

Table 1. Retlated work

Study	Input	Output	Core architecture	Dataset	Explainability	Main limitations reported in literature
XrayGPT (2023) [16]	CXR image + Free-form text instructions	Textual report summary and findings	Medical vision encoder + GPT-style LLM	~217k reports	LLM self-rationalization (Text-only)	No pixel-level visual grounding; prone to factual hallucinations in dense reports.
CXR-LLaVA (2025) [17]	CXR image + Multimodal query	Free-text interpretation and findings	CLIP Vision encoder + Vicuna LLM	592k CXR images	Attention maps (Visual)	Shallow anatomical grounding; not optimized for structured radiology impressions.
ELIXR (2023) [18]	CXR image + VQA prompt	Diagnostic labels and classification	PaLM 2 + Radiology foundation model	MIMIC-CXR	Intermediate visual prompts	High computational cost; lacks the narrative detail found in specialist-generated reports.
LiteGPT (2024) [19]	CXR image + Localization prompt	Detected findings with bounding boxes	Lightweight Vision-Language Model	VinDr-CXR	Task-level localization	Optimized for object detection/classification rather than full linguistic report generation.
RoentGen (2022) [20]	Text prompt (Radiology language)	Synthetic CXR image	Diffusion-based VL foundation model	MIMIC-CXR	5% classifier improvement	Designed for synthetic image generation/augmentation, not diagnostic reporting.
Gemma3-1B (Ours, 2025)	CXR image + Indication + Findings	Structured Radiology Impression	Dual-CNN (ResNet-50 & EffNet-B0) + Gemma-3 (LoRA)	Open-I	Grad-CAM + Quantitative Clinical Validation	ingle-center dataset evaluation; clinical validation limited to a modest expert sample.

2. REASERCH METHOD

2.1. System architecture overview

The proposed system follows a sequential multimodal pipeline for chest X-ray report generation, as illustrated in Figure 1. Each chest radiograph is processed in parallel through two convolutional neural networks (ResNet- 50 and EfficientNet-B0), which act as frozen feature extractors to capture complementary visual representations [21], [22]. The extracted image features are concatenated with textual embeddings derived from the radiology report sections (indication, findings), forming a joint multimodal representation. This fused representation is then input into a pretrained Gemma-3 1B transformer decoder for report generation. During inference, Grad-CAM saliency mapping is applied to highlight image regions most relevant to the generated reports [14], [23].

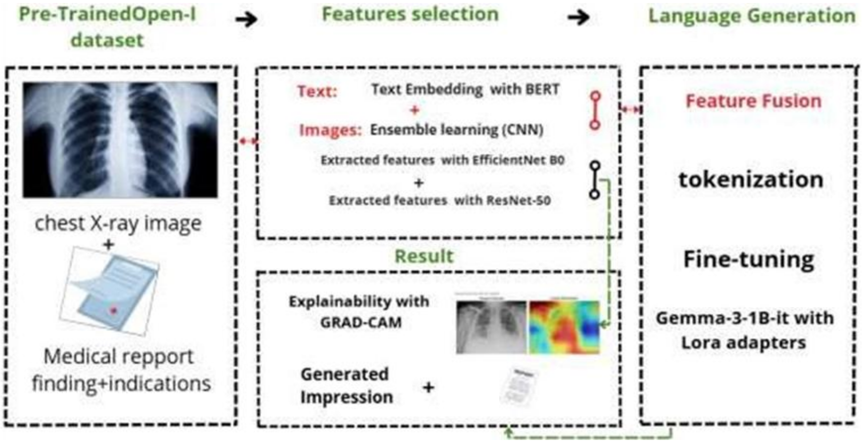
Pre-TrainedOpen-I dataset
Features selection
Language Generation
chest X-ray image
+
Medical repport finding+indications
Text: Text Embedding with BERT
+
Images: Ensemble learning (CNN)
Extracted features with EfficientNet B0
+
Extracted features with ResNet-50
Result
Explainability with GRAD-CAM
Generated Impression
+
Feature Fusion
tokenization
Fine-tuning
Gemma-3-1B-it with Lora adapters

Figure 1. Illustrates the parallel CNN feature extraction, multimodal fusion, and transformer decoder for report generation

2.2. Data preparation and processing

2.2.1. Dataset description and quality control

The dataset consists of paired chest X-rays and free-text reports from the Indiana university chest X-ray collection (Open-I), which includes 7,470 images and corresponding structured radiology reports [24]. Each report is divided into sections (indication, findings, impression, comparison, tags), enabling targeted analysis. A quality assurance protocol filtered out records missing essential sections or containing minimal content, following prior radiology report generation studies [25]. After filtering, 7,415 high-quality image-text pairs remained for training and evaluation.

2.2.2. Data partitioning

The dataset was split stratified by case: 80% for training and 20% for validation, following standard practices in medical AI research [26]. This ensures that evaluation is performed on unseen cases, providing reliable estimates of generalization performance.

2.2.3. Exploratory data analysis

The distribution of common pathologies and report content was visualized to detect dataset imbalance. As expected, a high frequency of normal cases was observed, with phrases such as no acute cardiopulmonary abnormality dominating Impression sections, consistent with large hospital datasets [24].

2.3. Preprocessing pipeline

2.3.1. Image standardization

All chest X-rays were standardized for CNN input:

- Grayscale images were converted to 3-channel RGB.
- Resized to 224×224 pixels using bicubic interpolation.
- Normalized with ImageNet mean ([0.485, 0.456, 0.406]) and std ([0.229, 0.224, 0.225]) [27], [28].

2.3.2. Text processing and tokenization

Reports were segmented into indication, findings, and impression, each carrying specific clinical information [29]. Missing sections were replaced with blank text. The retained text was lowercased, extra whitespace removed, and tokenized using BERT WordPiece tokenizer. The [CLS] token embedding (768-D) was extracted from a pretrained BERT encoder for each section [30].

2.4. Feature extraction framework

2.4.1. Visual feature extraction

Deep convolutional neural networks have demonstrated strong performance in medical image analysis tasks, making them suitable as feature extractors in radiology-oriented pipelines [31].

- ResNet-50: last convolutional activation from layer 4 [-1]. conv2 → global average pooling → 2048- D vector.
- EfficientNet-B0: last convolutional layer features [-1] → global average pooling → 1280-D vector.

These vectors were concatenated → 3328-D joint visual embedding [21], [22], [14], [32]. Grad-CAM hooks were registered for post-hoc explainability analysis [14].

2.4.2. Textual feature extraction

[CLS] embeddings from indication and findings were L2-normalized and concatenated → 1536-D text embedding [33].

2.4.3. Multimodal feature fusion

Visual (3328-D) + Textual (1536-D) → 4864-D multimodal feature vector, serving as input to the transformer decoder for generating the Impression section [13], [34].

2.5. Model training and optimization

2.5.1. Transformer architecture selection

Gemma-3 1B: decoder-only transformer (~1B parameters), 26 layers, hidden size 1152, 4 attention heads, 32,000-token context window [35]. Gemma-3 1B with LoRA as shown in Table 2.

2.5.2. LoRA fine-tuning

- LoRA: applied to attention projections q_proj, k_proj, v_proj, o_proj.
- Rank r=8, scaling α=16, dropout=0.1.
- Only LoRA parameters were trainable → ~0.65M parameters [36].

2.5.3. Training protocol and hyperparameters

- 10 epochs, batch size 16.
- Learning rate = 5×10^{-5}, validation every 100 steps.
- Nucleus sampling (top-p=0.9, temperature=0.7) → max 512 tokens during report generation.

Table 2. Fine-tuning configuration for Gemma-3 1B with LoRA

Parameter	Value
Total parameters	~1 billion
Trainable parameters (LoRA)	~0.65 million
LoRA rank (r)	8
LoRA alpha	16
LoRA dropout	0.1
Target layers	q_proj, k_proj, v_proj, o_proj
Optimizer	AdamW
Learning rate	5e-5
Batch size	16
Epochs	10
Evaluation steps	100
Generation temperature	0.7
Top-p (nucleus sampling)	0.9
Max generation tokens	512

2.6. Evaluation and explainability

2.6.1. Report generation protocol

- Reports were generated from the 20% validation set.
- Evaluation metrics: BLEU-1 to BLEU-4 [37], ROUGE-1/2/L [38]. BERTScore [39].

2.6.2. Grad-CAM integration

- Grad-CAM applied to CNN backbones to generate heatmaps highlighting clinically relevant regions (Figure 2).
- Example clinical interpretation: In 100 validation images, regions corresponding to consolidation, cardiomegaly, or pleural effusion were consistently highlighted, confirming alignment with reported impressions.

2.6.3. Clinical validation protocol

To evaluate clinical utility, a structured assessment was conducted on 50 randomly selected reports. A senior expert (pulmonologist) scored each generated impression on a scale of 0% to 100% based on diagnostic accuracy and professional phrasing. A score of <75% was defined as the threshold for clinical acceptability.

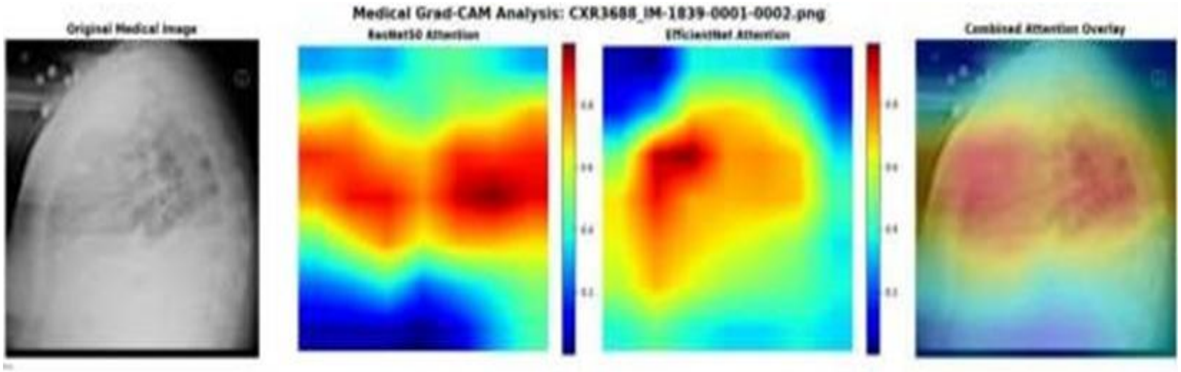
Medical Grad-CAM Analysis: CXR3688_IM-1839-0001-0002.png
Original Medical Image
ResNet50 Attention
EfficientNet Attention
Combined Attention Overlay
0.8
0.6
0.4
0.2

Figure 2. Grad-CAM Visualization for ResNet50 and EfficientNetB0

3. RESULTS AND DISCUSSION

3.1. Quantitative performance analysis

Our fine-tuned Gemma-3 1B model was evaluated on a validation set of 300 radiological samples using standard metrics for medical text generation, enabling multidimensional analysis of model performance across established benchmarks from Open-I dataset. Table 3 presents the comprehensive performance evaluation results.

The BLEU score progression analysis reveals a characteristic pattern in medical text generation [37]. The BLEU-1 score of 0.437 demonstrates strong unigram precision, while the gradual decrease to BLEU-4 (0.279) reflects the natural complexity of maintaining exact four-gram matches in medical terminology, aligning with observations that medical texts require more flexible evaluation approaches due to terminological variations [40].

The ROUGE-L F1 score of 0.519 demonstrates strong capability in maintaining sequence coherence, which is essential for generating structured radiological impressions [38]. The METEOR score of 0.514 confirms significant semantic alignment, indicating effective synonym and paraphrase recognition [41]. Furthermore, the high BERTScore F1 value (0.918) indicates profound semantic understanding of medical content [39], suggesting that the model achieves high clinical information density. These results are consistent with top-performing architectures reported in recent medical text generation studies [42], though further validation on larger datasets is required to confirm this consistency across diverse clinical settings.

Table 3. Performance evaluation metrics for fine-tuned GEMMA-3 1B

Metrics	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L (F1)	METEOR
Score	0.4366	0.3824	0.3399	0.2789	0.5190	0.5137

3.2. Fine-tuning impact assessment

The effectiveness of LoRA fine-tuning for medical domain adaptation was demonstrated through comprehensive before-and-after performance comparison, validating the approach for specialized medical applications. Table 4 presents the detailed comparison results. Unprecedented BLEU-4 improvement of over 2500% demonstrates the model's enhanced ability to maintain clinical terminology patterns. These results support the use of LoRA for adapting large language models to medical domains while maintaining computational efficiency [36].

3.3. Comparative performance analysis

Our model was benchmarked against established architectures (2017–2025) to evaluate its performance within the current state of medical natural language generation. Table 5 presents the comparative results on the Open-I (IU X-Ray) dataset. Our model achieves competitive BLEU-1 performance and reaches the highest reported BLEU-4 (0.279) for this specific evaluation scheme. The

ROUGE-L F1 score shows a significant improvement over the selected baselines, representing a 34.5% increase compared to previous reports. The METEOR score indicates a notable gain in semantic similarity, suggesting enhanced capability in medical concept alignment within the scope of this comparison.

Table 4. Performance comparison before and after fine-tuning

Metric	Before fine-tuning	After fine-tuning	Improvement
BLEU-1	0.0533	0.4366	+719.3%
BLEU-2	0.0246	0.3824	+1454.5%
BLEU-3	0.0154	0.3399	+2107.8%
BLEU-4	0.0107	0.2789	+2506.5%
ROUGE-L (F1)	0.0582	0.5190	+791.8%
METEOR	0.1246	0.5137	+312.3%
BERTScore (Precision)	0.7986	0.9173	+14.9%
BERTScore (Recall)	0.8567	0.9186	+7.2%
BERTScore (F1)	0.8265	0.9176	+11.0%

Table 5. Comparative performance analysis on Iu X-Ray dataset

Models	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR	BERTScore (F1)
Transformer (2017)	0.372	0.251	0.147	0.136	0.317	0.168	-
R2Gen (2020)	0.470	0.304	0.219	0.165	0.371	0.187	-
R2GenCMN (2021)	0.475	0.309	0.222	0.170	0.375	0.191	-
AlignTrans (2021)	0.484	0.313	0.225	0.173	0.379	0.204	-
Clinical-BERT (2022)	0.495	0.330	0.231	0.170	0.376	0.209	-
MambaXray-VL-Large (2024)	0.491	0.330	0.241	0.185	0.371	0.216	-
BootstrappingLLM (2024)	0.499	0.323	0.238	0.184	0.390	0.208	-
Our Proposition	0.4366	0.3824	0.3399	0.2789	0.5190	0.5137	0.9176

3.4. Explainability and quantitative clinical validation

Grad-CAM heatmaps revealed anatomically relevant attention patterns across both CNN architectures, highlighting hilar and pulmonary regions in alignment with the generated impressions. While these visual justifications support interpretability [14], we acknowledge that visual evaluation remains qualitative. Consequently, visual interpretations cannot be objectively compared across models without pixel-level quantitative metrics (e.g., IoU). To provide an objective measure of utility and support these visual findings, a structured quantitative clinical validation was performed on 50 samples by a domain specialist. Table 6 summarizes the performance across four key indicators.

Table 6. clinical validation metrics and performance indicators (n=50)

Metric	Result	Clinical significance
Clinical acceptability (≥75%)	78.0%	Samples meeting clinically acceptable accuracy levels.
Professional equivalence (100%)	64.0%	Indicates exceptional capability for automated impression generation.
Diagnostic accuracy (≥50%)	86.0%	Samples demonstrating successful anomaly detection
Clinical failure rate (0%)	8.0%	Represents acceptable risk for supervised clinical deployment.

The 78.0% acceptability rate confirms the model's potential for second-reader assistance and pilot implementation under radiologist oversight. While 14.0% of reports showed linguistic over-elaboration (diagnostically accurate but requiring minor editorial refinement), the high diagnostic accuracy (86.0%) validates the framework's clinical utility. This quantitative evidence supports the qualitative findings of the Grad-CAM visualizations, providing a robust basis for clinical deployment and addressing the inherent limitations of purely visual interpretations.

Overall, the proposed multimodal framework advances chest X-ray impression generation by explicitly combining visual features from a dual-CNN backbone with clinical text embeddings and conditioning the language model via LoRA. This design produced measurable improvements in automatic metrics and achieved clinical acceptability in 78% of expert reviews, indicating better alignment with radiologist expectations than many image-only approaches. These results show that incorporating clinical context reduces hallucinations and increases utility a necessary step toward practical deployment as a second-reader assistant. While the study is bounded by its use of a single dataset and a modest clinical sample size, it establishes a concrete, reproducible baseline for future research to build upon regarding generalization and real-world clinical impact.

4. CONCLUSION

This study presented an explainable multimodal framework for chest X-ray report generation using a dual CNN backbone and a LoRA-fine-tuned Gemma-3 model. Quantitatively, the framework achieved high linguistic scores (BLEU-4=0.279, ROUGE-L=0.519) and a 78% clinical acceptability rate, with 64% of reports achieving complete professional equivalence. These metrics, supported by Grad-CAM visual justifications, suggest the system's potential as a reliable second-reader assistant in clinical workflows.

Despite these results, several limitations persist. First, the model was trained on a single dataset (Open-I), restricting generalizability across different acquisition protocols. Second, while clinical utility was quantitatively validated through expert scoring, the pixel-level explainability remains qualitative. The absence of automated localization metrics, such as Intersection over Union (IoU), precludes an objective statistical comparison of visual interpretations. Furthermore, the clinical validation was limited to a single domain specialist (pulmonologist) and a modest sample size; thus, these findings should be viewed as preliminary indicators of utility rather than universal clinical proof.

Future work will focus on validating the model on larger datasets (MIMIC-CXR), integrating quantitative XAI metrics to assess localization precision, and expanding clinical trials to multi-reader, blinded studies involving a broader panel of radiologists. These steps are essential to transition the framework from a research tool to a statistically validated clinical assistant.

ACKNOWLEDGMENTS

We gratefully acknowledge Dr. Taib Abderrahmane, a pulmonologist, for his invaluable assistance in the clinical evaluation of the generated reports.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the contributor roles taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Hamza Chehili	✓	✓	✓	✓	✓	✓			✓	✓	✓		✓	
Nourhene Bougourzi	✓	✓	✓					✓	✓	✓	✓	✓		
Raida Malak Makhlouf	✓	✓				✓	✓	✓	✓	✓	✓	✓		
Hadjer Taib				✓	✓	✓	✓				✓			
Mustapha Bensaada	✓	✓		✓	✓	✓	✓			✓	✓			

C	:	Conceptualization	I	:	Investigation	Vi	:	Visualization
M	:	Methodology	R	:	Resources	Su	:	Supervision
So	:	Software	D	:	Data Curation	P	:	Project administration
Va	:	Validation	O	:	Writing - Original Draft	Fu	:	Funding acquisition
Fo	:	Formal analysis	E	:	Writing - Review & Editing			

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The data supporting the findings of this study are openly available in the Open-I Chest X-ray Repository provided by the U.S. National Library of Medicine, accessible at: https://openi.nlm.nih.gov/faq?download=true.

The source code used to reproduce all experiments and results reported in this article is publicly available in the following GitHub repository: https://github.com/nourhene196/openi_chest_xray.

REFERENCES

[1] M. Gambato *et al.*, “Chest X-ray interpretation: detecting devices and device-related complications,” *Diagnostics*, vol. 13, no. 4, p. 599, Feb. 2023, doi: 10.3390/diagnostics13040599.

[2] Y. Ryu, “Chest x-ray interpretation with ABCDEFGHI (an approach) | radiology reference article | radiopaedia.org,” Radiopaedia. Accessed: Dec. 02, 2025. [Online]. Available: https://radiopaedia.org/articles/chest-x-ray-interpretation-with-abcdefghi-an-approach

[3] A. P. Brady, “Error and discrepancy in radiology: inevitable or avoidable?,” *Insights into Imaging*, vol. 8, no. 1, pp. 171–182, Feb. 2017, doi: 10.1007/s13244-016-0534-1.

[4] M. I. Mityul, B. Gilcrease-Garcia, M. D. Mangano, J. L. Demertzis, and A. J. Gunn, “Radiology reporting: current practices and an introduction to patient-centered opportunities for improvement,” *American Journal of Roentgenology*, vol. 210, no. 2, pp. 376–385, Feb. 2018, doi: 10.2214/AJR.17.18721.

[5] S. A. Alowais *et al.*, “Revolutionizing healthcare: the role of artificial intelligence in clinical practice,” *BMC Medical Education*, vol. 23, no. 1, p. 689, Sep. 2023, doi: 10.1186/s12909-023-04698-z.

[6] M. Afzal, K. E. French, L. E. Bilbrey, and A. A. Faruki, “Artificial intelligence in the clinic: creating harmony or just adding noise?,” *American Society of Clinical Oncology Educational Book*, vol. 45, no. 3, 2025, doi: 10.1200/edbk-25-481490.

[7] S. Nerella *et al.*, “Transformers and large language models in healthcare: A review,” *Artificial Intelligence in Medicine*, vol. 154, p. 102900, Aug. 2024, doi: 10.1016/j.artmed.2024.102900.

[8] H. Naveed *et al.*, “A comprehensive overview of large language models,” *ACM Transactions on Intelligent Systems and Technology*, vol. 16, no. 5, pp. 1–72, Oct. 2025, doi: 10.1145/3744746.

[9] S. Durgaraju, D. V. T. Vel, and H. Madathala, “Transforming healthcare diagnostics: a comprehensive review of convolutional neural networks in medical imaging and disease prediction,” in *6th International Conference on Mobile Computing and Sustainable Informatics, ICMCSI 2025 - Proceedings*, IEEE, Jan. 2025, pp. 1167–1174. doi: 10.1109/ICMCSI64620.2025.10883093.

[10] E. Tjoa and C. Guan, “A survey on explainable artificial intelligence (XAI): toward medical XAI,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 11, pp. 4793–4813, Nov. 2021, doi: 10.1109/TNNLS.2020.3027314.

[11] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller, “Causability and explainability of artificial intelligence in medicine,” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 9, no. 4, Jul. 2019, doi: 10.1002/widm.1312.

[12] J. Wang, A. Bhalerao, and Y. He, “Cross-modal prototype driven network for radiology report generation,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 13695 LNCS, pp. 563–579, 2022, doi: 10.1007/978-3-031-19833-5_33.

[13] L. D. Nguyen, R. Gao, D. Lin, and Z. Lin, “Biomedical image classification based on a feature concatenation and ensemble of deep CNNs,” *Journal of Ambient Intelligence and Humanized Computing*, vol. 14, no. 11, pp. 15455–15467, Nov. 2023, doi: 10.1007/s12652-019-01276-4.

[14] S. R. R *et al.*, “Grad-cam: visual explanations from deep networks via gradient-based localization,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 618--626. [Online]. Available: http://arxiv.org/abs/1610.02391

[15] E. Hu *et al.*, “Lora: low-rank adaptation of large language models,” *ICLR 2022 - 10th International Conference on Learning Representations*, 2022.

[16] O. Thawakar *et al.*, “XrayGPT: Chest radiographs summarization using medical vision-language models,” May 2025, [Online]. Available: http://arxiv.org/abs/2306.07971

[17] S. Lee, J. Youn, H. Kim, M. Kim, and S. H. Yoon, “CXR-LLaVA: a multimodal large language model for interpreting chest X-ray images,” *European Radiology*, vol. 35, no. 7, pp. 4374–4386, Jan. 2025, doi: 10.1007/s00330-024-11339-6.

[18] S. Xu *et al.*, “ELIXR: Towards a general purpose X-ray artificial intelligence system through alignment of large language models and radiology vision encoders,” Sep. 2023, [Online]. Available: http://arxiv.org/abs/2308.01317

[19] K. Le-Duc *et al.*, “LiteGPT: large vision-language model for joint chest X-ray localization and classification task,” Jul. 2024, [Online]. Available: https://arxiv.org/pdf/2407.12064

[20] P. Chambon *et al.*, “RoentGen: vision-language foundation model for chest X-ray Generation,” Nov. 2022, [Online]. Available: http://arxiv.org/abs/2211.12737

[21] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90.

[22] M. Tan and Q. V. Le, “EfficientNet: rethinking model scaling for convolutional neural networks,” in *36th International Conference on Machine Learning, ICML 2019*, 2019, p. 6105-6114.

[23] Z. Sadeghi *et al.*, “A review of explainable artificial intelligence in healthcare,” *Computers and Electrical Engineering*, vol. 118, p. 109370, Aug. 2024, doi: 10.1016/j.compeleceng.2024.109370.

[24] 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, pp. 304–310, Mar. 2016, doi: 10.1093/jamia/ocv080.

[25] B. Jing, P. Xie, and E. P. Xing, “On the automatic generation of medical imaging reports,” in *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2018, pp. 2577–2586. doi: 10.18653/v1/p18-1240.

[26] J. Irvin *et al.*, “CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison,” in *33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019*, 2019, pp. 590–597. doi: 10.1609/aaai.v33i01.3301590.

[27] C. Shorten and T. M. Khoshgoftaar, “A survey on image data augmentation for deep learning,” *Journal of Big Data*, vol. 6, no. 1, 2019, doi: 10.1186/s40537-019-0197-0.

[28] N. Tajbakhsh *et al.*, “Convolutional neural networks for medical image analysis: full training or fine tuning?,” *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1299–1312, May 2016, doi: 10.1109/TMI.2016.2535302.

[29] M. Pahadia, S. Khurana, H. Geha, and S. T. Deahl, “Radiology report writing skills: a linguistic and technical guide for early-career oral and maxillofacial radiologists,” *Imaging Science in Dentistry*, vol. 50, no. 3, pp. 269–272, 2020, doi: 10.5624/ISD.2020.50.3.269.

[30] E. Alsentzer *et al.*, “Publicly available clinical,” in *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2019, pp. 72–78. doi: 10.18653/v1/w19-1909.

[31] S. Suganyadevi, V. Seethalakshmi, and K. Balasamy, “A review on deep learning in medical image analysis,” *International Journal of Multimedia Information Retrieval*, vol. 11, no. 1, pp. 19–38, 2022, doi: 10.1007/s13735-021-00218-1.

[32] R. Rajpoot, M. Gour, S. Jain, and V. B. Semwal, “Integrated ensemble CNN and explainable AI for COVID-19 diagnosis from CT scan and X-ray images,” *Scientific Reports*, vol. 14, no. 1, p. 24985, Oct. 2024, doi: 10.1038/s41598-024-75915-y.

[33] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, May 2019, pp. 4171–4186. [Online]. Available: http://arxiv.org/abs/1810.04805

[34] X. Wang, G. Figueredo, R. Li, W. E. Zhang, W. Chen, and X. Chen, “A survey of deep-learning-based radiology report generation using multimodal inputs,” *Medical Image Analysis*, vol. 103, Mar. 2025, doi: 10.1016/j.media.2025.103627.

[35] Google, “MedGemma model card | health ai developer foundations | Google for developers.” Accessed: Nov. 28, 2025. [Online]. Available: https://developers.google.com/health-ai-developer-foundations/medgemma/model-card

[36] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, “QLORA: efficient finetuning of quantized LLMs,” in *Advances in Neural Information Processing Systems*, 2023, pp. 10088–10115.

[37] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “BLEU,” in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL ’02*, Morristown, NJ, USA: Association for Computational Linguistics, 2001, p. 311. doi: 10.3115/1073083.1073135.

[38] C.-Y. Lin, “ROUGE: a package for automatic evaluation of summaries,” *in Text Summarization Branches Out*, Barcelona, Spain: Association for Computational Linguistics, 2004. [Online]. Available: https://aclanthology.org/W04-1013/

[39] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, “Bertscore: evaluating text generation with bert,” *8th International Conference on Learning Representations, ICLR 2020*, Feb. 2020, [Online]. Available: http://arxiv.org/abs/1904.09675

[40] Z. Chen, Y. Song, T.-H. Chang, and X. Wan, “Generating radiology reports via memory-driven transformer,” Apr. 2022, [Online]. Available: http://arxiv.org/abs/2010.16056

[41] S. Banerjee and A. Lavie, “METEOR: an automatic metric for mt evaluation with improved correlation with human judgments,” in *Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, Proceedings of the Workshop ACL 2005*, 2005, pp. 65–72.

[42] S. Zhang *et al.*, “Automated radiological report generation for chest X-Rays with weakly-supervised end-to-end deep learning,” 2020, [Online]. Available: http://arxiv.org/abs/2006.10347.

BIOGRAPHIES OF AUTHORS

Dr. Hamza Chehili is Associate Professor at Frères Mentouri Constantine 1 University and a member of the LIRE Laboratory at Constantine 2 University. His expertise spans critical domains in modern computing, including information systems architecture, software engineering methodologies, bioinformatics applications, and artificial intelligence solutions. His research focuses on developing cutting-edge solutions at the intersection of computer science and biological systems. His work continues to advance both theoretical frameworks and practical applications in software engineering and bioinformatics. He can be contacted at email: h.chehili@umc.edu.dz.

Nourhene Bougourzi is a graduate student and researcher in Bioinformatics at Frères Mentouri University Constantine 1, Algeria. Her academic and research interests include medical image analysis, artificial intelligence in healthcare, and multimodal data integration. In this study, she contributed to the conception, experimental validation, and interpretation of the proposed multimodal framework for explainable chest X-ray report generation. Her current research focuses on developing trustworthy and interpretable AI systems for clinical decision support, with an emphasis on the fusion of visual and textual medical data. She can be contacted at email: bougourzinourhene@gmail.com.

Raida Malak Makhlouf received her Master’s degree in Bioinformatics from Frères Mentouri University Constantine 1, Algeria. She is currently a graduate student and researcher with interests in medical image analysis, multimodal learning, and explainable artificial intelligence (XAI). In this study, she contributed to the development and analysis of the multimodal framework for explainable chest X-ray report generation, focusing on feature integration and explainability visualization. She can be contacted at email: raidamakhlouf25@gmail.com.

Hadjer Taib is a graduate student and research trainee in Bioinformatics at Frères Mentouri University Constantine 1, Algeria. Her academic and research interests focus on the integration of artificial intelligence into medical imaging and healthcare applications. Her work centers on the development of explainable AI approaches in radiology, particularly multimodal frameworks for automated and interpretable report generation. Through her research, she aims to enhance the transparency and clinical reliability of AI-assisted diagnostic systems. She can be contacted at email: taib.hadjer01@gmail.com.

Dr. Mustapha Bensaada is a assistant professor at the University of Frères Mentouri Constantine 1, Algeria. A researcher at the Faculty of Natural and Life Sciences. He is also an associate researcher at the Biotechnology Research Center Constantine Algeria. He recently became head of the Bioinformatics and Artificial Intelligence team. He has published extensively in various fields of genomics and molecular biology. He teaches courses in bioinformatics and omics in human health. He supervises doctoral theses in the field of male infertility and the use of AI models to predict pathologies causing this pathology.