This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

NaijaCXR-VLM: A Multi-task Vision-Language Model for Chest X-ray Analysis in Low-resource Settings

Aminu Musa[1,2], Rajesh Prasad[1,3], and Monica Hernandez[4,5]

[1] Department of Computer Science, African University of Science and Technology, Abuja, Nigeria
[2] Department of Computer Science, Federal University Dutse, Nigeria
[3] Department of Computer Engineering, Ghaziabad College of Engineering, Ghaziabad, India
[4] Department of Computer Science, University of Zaragoza, Zaragoza, Spain
[5] Aragon Institute on Engineering Research (I3A), University of Zaragoza, 50018 Zaragoza, Spain
musa.aminu@fud.edu.ng

Abstract. Large vision–language models (LVLMs) have demonstrated strong performance in chest X-ray analysis and automated radiology report generation, but often exhibit racial bias, leading to lower accuracy on data from low- and middle-income countries (LMICs). To address these challenges, we evaluated existing Med-VLMs on Naija-CXR, a newly curated dataset from a tertiary hospital in Nigeria, and on the MIMIC-CXR dataset from the United States. Existing MedVLMs show population bias. To mitigate poor performance due to population-domain shifts, we introduce NaijaCXR-VLM, a lightweight, multitask framework that combines supervised adversarial adaptation and parameter-efficient fine-tuning for pathology classification and report generation. Built on MedGemma-4B, our model employs gradient-reversal and supervised alignment modules to learn domain-invariant representations, thereby mitigating population bias. This method enhances cross-domain generalization and achieves 80.23% accuracy, 50.94 BLEU-1, 57.82 RadGraph, and 65.22% AUROC on Naija-CXR. NaijaCXR-VLM outperforms all baselines models on report generation performance and achieves highest F1-score, precision, and recall for pathology classification, highlighting the role of domain adaptation in mitigating population bias.

Keywords: Bias mitigation, Chest X-ray, Domain adaptation, Fairness, Medical imaging, Vision-language models.

1 Introduction

The increasing number of preventable deaths in low- and middle-income countries (LMICs) can largely be attributed to delayed diagnoses, resulting from a severe shortage of radiologists, inadequate infrastructure, and healthcare resources [1]. In Nigeria, the ratio of radiologists to the population is approximately

1:10,000 in urban areas and drops drastically to 1:100,000 in rural regions [2]. This problem is particularly acute in the northern region of the country, which has the lowest physician-to-population ratio nationwide [3]. Addressing this critical challenge is crucial for advancing equitable healthcare delivery in low-resource settings.

The integration of AI-assisted diagnostic systems in radiology has enabled the automated analysis of medical images and the generation of preliminary reports and diagnoses. This integration offers a transformative opportunity to aid diagnosis, particularly in underserved areas. Such systems can substantially reduce radiologists' workloads, enhance diagnostic efficiency, and reduce the likelihood of errors and delays [4].

The remarkable success of large language models (LLMs) has spurred the development of Vision–language models (VLMs), which integrate computer vision and natural language processing (NLP) to analyze visual and textual information and generate multimodal outputs jointly [5]. The performance of VLMs on general-purpose tasks has motivated their adaptation to specialized domains, particularly in medical imaging, to enable medical imaging analysis and computer-aided diagnosis (CAD), leading to the development of medical VLMs such as Med-Gemma [6], Qwen-VL2 [7], and LlaVA-Med [8], which are trained on large-scale medical datasets to serve as foundational models for AI-assisted diagnostics.

Specifically, in chest X-ray (CXR) analysis, several vision–language models have been introduced recently to enhance clinical understanding and automated report generation. Models such as RADVLM [9], CheXagent [10], and Meta-CXR [11] represent notable advances in medical VLM development. Despite the impressive Performance of these models, their performance tends to drop when evaluated on an unseen dataset from a different population distribution [12]. While Musa et al [12] investigate supervised domain adaptation for chest X-ray classification, our work addresses the substantially different task of free-text radiology report generation using a Vision-Language Model, and critically demonstrates that adversarial domain adaptation strategies effective for classification can degrade generative VLM performance, motivating native fine-tuning as a more reliable approach for this setting.

Deploying VLMs as CAD tools in clinical workflows poses challenges, including fairness concerns, bias, and model hallucinations. Most VLMs are trained on data from developed nations, such as MIMIC-CXR [13] and CheXpert [14], which underrepresent the LMIC population, thereby limiting generalizability and increasing the risk of algorithmic bias. This often leads to underperformance on non-representative datasets, like African cohorts, and on diseases prevalent in LMIC, such as tuberculosis and pneumonia, which are less represented in public datasets.

Therefore, we introduce NaijaCXR-VLM, a multi-task chest X-ray analysis model designed for both disease classification and radiology report generation, with an emphasis on equity and fairness. Specifically, the proposed NaijaCXR-VLM leverages Med-Gemma as a foundational backbone model, with Parameter-

Efficient Fine-Tuning (PEFT) [15] for lightweight deployment. The framework integrates a supervised adversarial domain adaptation (ADA) module, which employs a gradient reversal layer and a domain discriminator to align latent feature distributions between Western source cohorts (MIMIC-CXR) and the Nigerian target population, effectively mitigating cross-population domain shift (CPDS). To preserve diagnostic accuracy during alignment, we introduce a supervised class-wise centroid alignment loss that explicitly anchors the source and target representations sharing the same pathology label. The model is jointly optimized through a unified multi-task objective that couples multi-class classification, autoregressive report generation, adversarial domain confusion, and semantic alignment.

Unlike the work of [16] on an unsupervised domain adaptation framework, which uses gradient reversal to align feature distributions without supervision, our approach employs supervised adversarial adaptation with limited target-domain labels. This enables preservation of clinically meaningful features and better handles severe cross-population shifts. By combining supervised alignment, multitask learning, and parameter-efficient fine-tuning, our method is designed to achieve equitable, robust performance in low-resource clinical settings, ensuring clinically relevant outcomes.

2 Methodology

2.1 Datasets

Naija-CXR Dataset: Naija-CXR is a newly curated chest X-ray dataset collected at Rasheed Shekoni Federal University Teaching Hospital (RSFUTH), Dutse, Nigeria. The dataset contains 2,000 de-identified frontal chest radiographs and corresponding radiology reports retrospectively obtained from routine clinical practice. The dataset covers regionally prevalent diseases, including tuberculosis, pneumonia, pleural effusion, Heart failure, cardiomegaly, edema, and other lung consolidations. Images were acquired using varying equipment commonly found in low-resource hospitals, introducing distributional characteristics distinct from those in Western datasets.

The dataset comprises adult patients, predominantly aged 25–50 years, with a sex distribution of 78% female and 22% male. Pediatric studies were excluded to maintain cohort consistency. We also excluded radiographs containing objects that could introduce confounding visual artifacts. All images and associated radiology reports were fully anonymized. Two consultant radiologists performed image labeling and report verification. Each radiologist independently reviewed the images and reports, and disagreements were resolved by consensus to produce clinically reliable annotations.

MIMIC-CXR. MIMIC-CXR [13] is a large-scale publicly available dataset containing 377,110 chest X-ray (CXR) images and 227,827 corresponding free-text radiology reports collected from Beth Israel Deaconess Medical Center in the

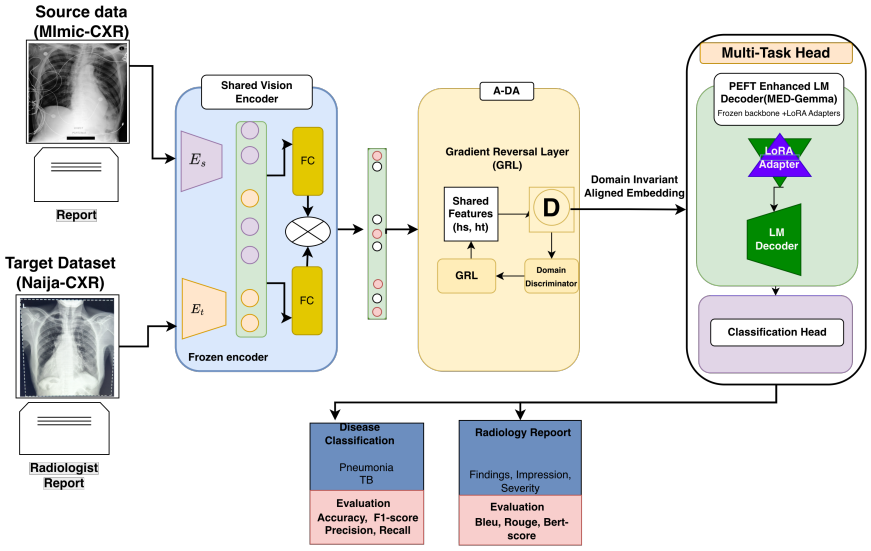

Fig. 1: Proposed NaijaCXR-VLM Architecture: A PEFT-enhanced Med-Gemma vision encoder extracts visual features, which are fused and passed to an adversarial domain adaptation module. The aligned representations are then fed into multi-task learning with a classification head and a LoRA-enabled Med-Gemma language decoder for radiology report generation.

United States. The dataset includes frontal and lateral views, multiple acquisition devices, and extensive coverage of cardiopulmonary conditions. We screened the images and excluded lateral views (122,500), since NaijaCXR contains only Frontal views. Finally, we selected 30,327 image-report pairs.

2.2 Proposed NaijaCXR-VLM

We build NaijaCXR-VLM upon Med-Gemma by localizing it using supervised adversarial domain adaptation and Parameter-efficient fine-tuning (PEFT). The architecture consists of a vision encoder for CXR images, a text encoder for radiology reports, a multimodal fusion module, a classification head for pathology prediction, and an autoregressive decoder for report generation.

The Proposed model framework integrates multimodal representation learning for cross-population chest X-ray analysis. It processes source-domain (MIMIC-CXR) and target-domain (NaijaCXR) images to produce clinically meaningful, domain-robust representations, making it suitable for deployment in low-resource settings. The model architecture is further illustrated in Fig.1.

The Source-domain (E_s) and target-domain (E_t) features are extracted via a shared, frozen SigLIP vision encoder. The resulting token embeddings are projected through fully connected layers and fused into a unified representation (h_s, h_t). The fused features are jointly optimized by an Adversarial Do-

main Adaptation (A-DA) module, which integrates a Gradient Reversal Layer (GRL), a domain discriminator ($D(\cdot)$), and supervised class alignment to enforce domain-invariant embeddings.

2.3 Model Formulation

We formalize NaijaCXR-VLM as a multi-task vision–language model trained jointly for (i) multi-class chest X-ray classification and (ii) radiology report generation, while mitigating domain shift via supervised adversarial domain adaptation.

Let an input chest X-ray be denoted by $x \in \mathcal{X}$, with associated label $y \in \{1, \ldots, K\}$ and report sequence $r = (r_1, \ldots, r_T)$. We denote the source-domain dataset (MIMIC-CXR) as:

$$\mathcal{D}_s = \{(x_i^s, r_i^s, y_i^s)\} \tag{1}$$

and the target-domain dataset (NaijaCXR) as:

$$\mathcal{D}_t = \{(x_j^t, r_j^t, y_j^t)\} \tag{2}$$

Our objective is to learn a vision-language model f_θ that achieves high performance on both domains by learning domain-invariant yet label-discriminative representations. **Vision Encoder.** The PEFT-enhanced Med-Gemma vision encoder produces visual token embeddings $v = f_{\theta_V, \Delta_\phi}(x) \in \mathbb{R}^{L \times d_v}$, where θ_V denotes frozen backbone parameters and Δ_ϕ are trainable PEFT adapters. The pooled visual feature is $h = \text{Pool}(v) \in \mathbb{R}^d$.

Adversarial Domain Adaptation. Let $d \in \{0, 1\}$ denote the domain label ($0 =$ source, $1 =$ target). Features are passed through a Gradient Reversal Layer (GRL) before the domain discriminator D_ψ: $\hat{d} = D_\psi(\text{GRL}_\lambda(h)) \in [0, 1]$. The adversarial loss is:

$$\mathcal{L}_{adv} = -\frac{1}{B} \sum_{i=1}^{B} \Big[d_i \log(\hat{d}_i) + (1 - d_i) \log(1 - \hat{d}_i) \Big] \tag{3}$$

Supervised Feature Alignment. To minimize the distance between same-class samples across domains, we define per-class source and target centroids $\mu_s^{(k)}$ and $\mu_t^{(k)}$ over minibatch subsets $B_s^{(k)}, B_t^{(k)}$, and compute the alignment loss:

$$\mathcal{L}_{align} = \frac{1}{K} \sum_{k=1}^{K} \left\| \mu_s^{(k)} - \mu_t^{(k)} \right\|_2^2 \tag{4}$$

where $\mu_s^{(k)} = \frac{1}{|B_s^{(k)}|} \sum_{i \in B_s^{(k)}} h_i^s$ and $\mu_t^{(k)}$ is defined analogously for the target domain. Classification uses standard cross-entropy loss $\mathcal{L}_{cls}$, and report generation is trained with an autoregressive language modeling loss $\mathcal{L}_{rg}$ over the conditional likelihood $p(r \mid v) = \prod_{t=1}^{T} p(r_t \mid r_{<t}, v; \theta_L, \Delta_\phi)$.

Total Training Objective. The final multi-task objective combines all losses:

$$\mathcal{L}_{total} = \mathcal{L}_{cls} + \alpha\mathcal{L}_{rg} + \beta\mathcal{L}_{adv} + \gamma\mathcal{L}_{align} \tag{5}$$

with weights $\alpha, \beta, \gamma \geq 0$. Trainable parameters include the PEFT adapters Δ_ϕ, classifier weights W, b, domain discriminator ψ, and LM head components; backbone parameters θ_V, θ_L remain frozen.

2.4 Experimental Setup

For fine-tuning the proposed model, both datasets were split into 70/15/15 train–validation–test sets. Cross-domain performance was achieved using the proposed domain adaptation approach. Consistent text and image preprocessing across datasets minimized confounding variables, contributing to reliable, high model performance.

The models were fine-tuned using AdamW with a learning rate of $1.0e^{-4}$, a batch size of 16, and early stopping based on validation loss. We used LoRA with parameters (lora alpha=32, rank=16, dropout=0.05, max new tokens=200), and a report generation decoder with temperature=0.2 and top_p=0.9, for lightweight deployment. All experiments were conducted on NVIDIA A100 GPUs with 80GB RAM. Training was performed for 5-10 epochs, depending on convergence.

Model performance was assessed using Accuracy, AUC, F1-score, precision, and recall for the classification task. For the report generation task, we used NLP metrics such as BLEU [17], ROUGE-L [18], and BERTscore [19], while for clinical accuracy and correctness, we used Radgraph-F1 [20] score and expert radiologist rating (on a 5-point scale).

2.5 Data Availability

The study releases all the source codes, model weights, configuration files, and data processing pipelines in the GitHub repository: url https://github.com/Elamin/NaijaCXR-VLM. The repository includes everything needed to reproduce the experiments exactly, including training scripts and documentation. The curated Naija-CXR dataset and related resources are made openly available at `https://zenodo.org/records/21781842.`

3 Results and Discussion

This section presents quantitative and qualitative evaluations of NaijaCXR-VLM for both chest radiograph classification and radiology report generation tasks. We evaluate the model's performance on the source dataset (MIMIC-CXR) and the target dataset.

3.1 Baseline VLM Performance

To establish baseline performance, we evaluate existing vision-language models (VLMs) on both MIMIC-CXR and Naija-CXR under two settings: zero-shot and fine-tuning. In the zero-shot setting, models are evaluated without task-specific adaptation to assess their cross-domain generalization. Whereas in the fine-tuning setting, each model is fine-tuned using the Naija-CXR training split to evaluate the benefits of limited target-domain supervision. The results are summarized in Tables 1 and 2.

Table 1: Zero-shot performance of medical VLMs on MIMIC-CXR and Naija-CXR.

	MIMIC-CXR				**Naija-CXR**			
Model	Bleu-1	Rouge-L	BertScore	RG-F1	Bleu-1	Rouge-L	BERTScore	RG-F1
Qwen3-VL [7]	7.3	14.0	42.0	21.0	3.3	16.0	40.6	16.3
Med-Gemma 4B [6]	17.5	20.7	47.7	29.5	2.6	31.3	84.5	13.9
RadVLM [9]	7.0	24.3	51.0	17.7	4.7	22.3	81.4	10.5
LLaVA-Med [8]	11.3	12.6	30.1	4.7	1.6	21.1	84.3	5.3
CheXagent [10]	14.3	22.9	38.1	20.9	1.9	19.2	85.8	7.4

Table 1 reports the BLEU, ROUGE, RadGraph F1, and BERTScore of all baseline models. All models exhibit substantially lower performance on Naija-CXR than on MIMIC-CXR on Bleu-1 and RadGraph F1 metrics, reflecting a significant cross-population domain shift driven by differences in imaging characteristics and radiology reporting styles. Consequently, the high BERTScore observed on Naija-CXR is because BERTScore relies on contextual embedding similarity; it may assign high scores to semantically similar but clinically incomplete reports. This limitation reinforces the need to complement semantic metrics with clinically oriented measures such as RadGraph F1.

Table 2 summarizes the fine-tuned performance of all baseline models and the proposed NaijaCXR-VLM on Naija-CXR, covering both report generation and classification. Among the baselines, RadVLM achieved the highest Accuracy and a comparable Macro-AUROC for classification, while NaijaCXR-VLM achieved the strongest overall report generation performance. The fine-tuned approach significantly enhances performance for nearly all models, demonstrating that incorporating the local dataset improves performance and reduces bias.

To evaluate the contribution of each component, we conducted an ablation study across four configurations: (1) parameter-efficient fine-tuning (PEFT) alone, (2) PEFT with an adversarial gradient reversal layer (GRL), (3) PEFT + GRL with an added domain discriminator, and (4) the full *NaijaCXR-VLM* model. As shown in Table 3, adding PEFT or adversarial components in isolation degrades performance relative to the baseline, indicating that domain adaptation alone, without the complementary components, is insufficient and can destabilize training. However, the full model, which integrates all components jointly, yields consistent gains across all metrics, improving F1-Score, AUROC, and RadGraph F1 over the baseline.

Table 2: Performance comparison of medical VLMs and the proposed NaijaCXR-VLM on Naija-CXR, for radiology report generation and multi-class pathology classification (fine-tuned on the Naija-CXR training split).

Models	Report Generation				Classification (%)				
	BLEU-1	ROUGE-L	BERTScore	RG-F1	Acc.	F1	Prec.	Rec.	AUROC
Qwen3-VL	49.62	66.22	93.09	50.39	66.83	61.24	44.14	55.86	50.05
Med-Gemma 4B	10.87	15.36	82.19	13.54	65.50	53.72	40.69	62.99	58.68
RadVLM	35.93	59.44	89.34	45.19	**82.21**	57.74	44.68	43.02	65.21
LLaVA-Med	0.88	12.01	82.29	10.11	59.96	18.10	22.84	34.19	53.92
CheXagent	33.08	26.57	83.78	17.95	68.31	46.34	57.66	37.71	53.13
NaijaCXR-VLM (Ours)	**50.94**	**68.95**	**94.08**	**57.82**	80.23	**72.34**	**57.67**	**67.91**	**65.22**

Table 3: Ablation study on NaijaCXR. Each component is added cumulatively to highlight its contribution.

Model Variant	F1-Score	AUROC	RadGraph F1
Baseline Med-Gemma (no PEFT)	66.85	58.68	13.54
+ PEFT Only	50.88	52.87	7.78
+ PEFT + GRL (Adversarial Only)	47.89	54.99	10.21
+ PEFT + GRL + Domain Discriminator + Alignment Loss (no joint MTL)	53.72	58.28	13.54
NaijaCXR-VLM (Full Model)	**72.34**	**65.22**	**57.82**

3.2 Discussion

Zero-shot Baseline Performance. Table 1 summarizes the zero-shot performance of existing medical VLMs on MIMIC-CXR and Naija-CXR. Across all models, BLEU-1 and RadGraph F1 decline substantially on Naija-CXR, indicating poor lexical precision and reduced preservation of clinically relevant information under cross-population domain shift. In contrast, several models achieve higher ROUGE-L and higher BERTScore on Naija-CXR. This apparent discrepancy arises because BERTScore and, to a lesser extent, ROUGE-L primarily measure semantic or contextual similarity and can be inflated by repetitive report templates or generic clinical language, even when key diagnostic findings are omitted or incorrect. Consequently, we place greater emphasis on BLEU-1 and RadGraph F1, which better reflect clinical correctness. Overall, the zero-shot results demonstrate limited generalization of existing VLMs to the Nigerian clinical domain.

Proposed NaijaCXR-VLM. As shown in Table 2, NaijaCXR-VLM substantially improves radiology report generation on Naija-CXR, achieving a BLEU-1 of 50.94%, ROUGE-L of 68.95%, BERTScore of 94.08%, and RadGraph F1 of 57.82%, outperforming all baseline models. For multi-class classification, the proposed model achieves an accuracy of 80.23%, an F1-score of 72.34%, and an AUC of 65.22%. While RadVLM attains slightly higher accuracy among the fine-tuned baselines, NaijaCXR-VLM provides the best overall report generation performance while maintaining competitive classification performance through a unified multitask framework. These improvements suggest that adversarial domain adaptation and multitask learning effectively reduce cross-population domain shift and improve clinical report quality.

4 Conclusion

We presented NaijaCXR-VLM, a lightweight domain-adaptive vision–language framework designed to improve robustness under population shift in chest X-ray analysis. Compared to a baseline MedGemma model trained on MIMIC-CXR, the proposed approach achieved improved performance on Naija-CXR, including gains in classification accuracy (80.23%), BLEU-1 (50.94), RadGraph (57.82), and AUROC (65.22%). The combination of supervised adversarial domain adaptation and parameter-efficient fine-tuning enables lightweight adaptation suitable for resource-constrained clinical environments. The study's limitations include a small dataset and a single-institution setting. Future work will also incorporate an expanded set of visualizations, including a more comprehensive pipeline diagram and qualitative report-generation examples, to further improve reader accessibility.

Ethical Approval

The collection and curation of the Naija-CXR dataset were approved by the Research Ethics Committee of Rasheed Shekoni Federal University Teaching Hospital, Dutse, Nigeria (approval no. 226/V.II). All radiographs and reports were obtained retrospectively from routine clinical practice and fully de-identified before use; the approving committee waived the requirement for informed consent given the retrospective, de-identified nature of the data. We used the MIMIC-CXR dataset under a PhysioNet data use agreement, in accordance with its terms.

Disclosure of Interests

The authors have no competing interests to declare that are relevant to the content of this article.

References

1. A. Adeolu, O. Arowolo, O. Alatise, S. Osasan, W. Odesanmi, and E. Omoniyi, "Pattern of death in a nigerian teaching hospital; 3-decade analysis," *African health sciences*, vol. 10, no. 3, 2010.
2. B. M. Idowu and T. A. Okedere, "Diagnostic radiology in nigeria: a country report," *Journal of Global Radiology*, vol. 6, no. 1, 2020.
3. W. Igoniye and C. Njeh, "Navigating radiological challenges in nigeria: A comprehensive review," *MEDICAL PHYSICS*, vol. 12, no. 2, 2024.
4. C. Mello-Thoms and C. A. Mello, "Clinical applications of artificial intelligence in radiology," *The British Journal of Radiology*, vol. 96, no. 1150, p. 20221031, 2023.
5. F. Bordes, R. Y. Pang, A. Ajay, A. C. Li, A. Bardes, S. Petryk, O. Mañas, Z. Lin, A. Mahmoud, B. Jayaraman *et al.*, "An introduction to vision-language modeling," *arXiv preprint arXiv:2405.17247*, 2024.

6. A. Sellergren, S. Kazemzadeh, T. Jaroensri, A. Kiraly, M. Traverse, T. Kohlberger, S. Xu, F. Jamil, C. Hughes, C. Lau *et al.*, "Medgemma technical report," *arXiv preprint arXiv:2507.05201*, 2025.
7. Y. Yu, C. Shi, J. Tang, and S. Zheng, "Qwen-vl2 model with neftune technique for medical report generation," in *2025 4th International Symposium on Computer Applications and Information Technology (ISCAIT)*. IEEE, 2025, pp. 165–168.
8. C. Li, C. Wong, S. Zhang, N. Usuyama, H. Liu, J. Yang, T. Naumann, H. Poon, and J. Gao, "Llava-med: Training a large language-and-vision assistant for biomedicine in one day," *Advances in Neural Information Processing Systems*, vol. 36, pp. 28 541–28 564, 2023.
9. N. Deperrois, H. Matsuo, S. Ruipérez-Campillo, M. Vandenhirtz, S. Laguna, A. Ryser, K. Fujimoto, M. Nishio, T. M. Sutter, J. E. Vogt *et al.*, "Radvlm: A multitask conversational vision-language model for radiology," *arXiv preprint arXiv:2502.03333*, 2025.
10. Z. Chen, M. Varma, J.-B. Delbrouck, M. Paschali, L. Blankemeier, D. Van Veen, J. M. J. Valanarasu, A. Youssef, J. P. Cohen, E. P. Reis *et al.*, "Chexagent: Towards a foundation model for chest x-ray interpretation," *arXiv preprint arXiv:2401.12208*, 2024.
11. D. Edirisinghe, W. Nimalsiri, M. Hennayake, D. Meedeniya, and G. Lim, "Chest x-ray report generation using abnormality guided vision language model," *IEEE Access*, 2025.
12. A. Musa, M. Ibrahim Adamu, H. A. Kakudi, M. Hernandez, and Y. Lawal, "Analyzing cross-population domain shift in chest x-ray image classification and mitigating the gap with deep supervised domain adaptation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2024, pp. 585–595.
13. A. E. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, R. G. Mark, and S. Horng, "Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports," *Scientific data*, vol. 6, no. 1, p. 317, 2019.
14. J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya *et al.*, "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 590–597.
15. N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, "Parameter-efficient transfer learning for nlp," in *International conference on machine learning*. PMLR, 2019, pp. 2790–2799.
16. Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *International conference on machine learning*. PMLR, 2015, pp. 1180–1189.
17. M. Post, "A call for clarity in reporting bleu scores," in *Proceedings of the third conference on machine translation: Research papers*, 2018, pp. 186–191.
18. C.-Y. Lin and F. J. Och, "Looking for a few good metrics: Rouge and its evaluation," in *Ntcir workshop*, 2004, pp. 1–8.
19. T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," *arXiv preprint arXiv:1904.09675*, 2019.
20. S. Jain, A. Agrawal, A. Saporta, S. Q. Truong, D. N. Duong, T. Bui, P. Chambon, Y. Zhang, M. P. Lungren, A. Y. Ng *et al.*, "Radgraph: Extracting clinical entities and relations from radiology reports," *arXiv preprint arXiv:2106.14463*, 2021.