Comparative Development of BioMedCLIP for Enhanced Biomedical Data Integration

Praveen Pandey
Department of Artificial Intelligence, Amity School of Engineering & Technology, Amity University, Noida, India
praveenpandey.ai@gmail.com

Hiyaa Malik
Department of Artificial Intelligence, Amity School of Engineering & Technology, Amity University, Noida, India
hiyaamalik205@gmail.com

Sofia Singh
Department of Artificial Intelligence, Amity School of Engineering & Technology, Amity University, Noida, India
ssingh5@amity.edu (corresponding author)

Dipti Theng
Department of Computer Science & Engineering, Symbiosis Institute of Technology Pune, Symbiosis International (Deemed University), Pune, India
deepti.theng@gmail.com

Urvashi Agrawal
Department of Electronics & Telecommunication Engineering, Jhulelal Institute of Technology, Nagpur, India
urvashi.agrawal2000@gmail.com

Raj Kumar
Department of Computer Science & Engineering, Indian Institute of Technology, Patna, Bihar, India
raj_25s09res57@iitp.ac.in

Sanjay Balwani
Department of Electronics & Telecommunication Engineering, Jhulelal Institute of Technology, Nagpur, India
sanjaybalwani31@gmail.com

Anoop Kumar Shukla
Amity University, Noida, India
akshukla1@amity.edu

Received: 30 September 2025 | Revised: 2 November 2025 | Accepted: 12 November 2025

 DOI: https://doi.org/10.48084/etasr.15242

ABSTRACT

Advancements in multimodal learning in Artificial Intelligence (AI) have led to the proposal of many foundational models in the area of biomedical AI. However, leveraging these models for specialized clinical use requires a satisfactory level of fine-tuning and validation. BioMedCLIP is one such state-of-the-art model. This paper presents a comparative study that focuses on adapting BioMedCLIP and evaluating its

efficiency across various datasets and fine-tuning approaches. Several fine-tuning methods were explored, and the model was implemented on two large and challenging datasets. The first dataset was the large-scale National Institutes of Health (NIH) Chest X-ray collection for multi-label disease classification; the second dataset was HAM10000, a large collection of multi-source dermatoscopic images of skin lesions. The primary objective was to assess different fine-tuning strategies and develop a specialized model with enhanced capabilities in integrating images and textual data. Using three different approaches, the study compares BioMedCLIP across the NIH Chest X-ray and HAM10000 datasets, demonstrating improved text–image modality integration, 55.6% macro accuracy, and reduced overfitting. This work validates the effectiveness of domain-specific adaptation and establishes a powerful, fine-tuned model ready for deployment in advanced healthcare applications.

Keywords-BioMedCLIP; finetuning; multimodal Artificial Intelligence (AI); Contrastive Language–Image Pretraining (CLIP); fusion; cross-attention transformer

I. INTRODUCTION

Undoubtedly, the healthcare and biomedical sectors have seen rapid advancements, especially in the domain of Artificial Intelligence (AI). Multimodal AI is one such area. The emergence of multimodal learning has influenced computational models designed to aid in diagnosis and clinical decision making. These models operate by integrating information from various input modalities. More specifically, foundational models trained on large-scale datasets show good performance in downstream tasks. However, using them directly for specialized domains remains questionable due to the unique complexity, variability, and domain-specific nature of medical data.

To bridge this gap, domain-adapted models such as BioMedCLIP have been developed, which offer advanced multimodal capabilities, especially for biomedical use cases. BioMedCLIP is a biomedical vision–language foundation model pretrained on PMC-15M, a dataset of 15 million figure–caption pairs extracted from biomedical research articles. It uses PubMedBERT as the text encoder and a vision transformer as the image encoder. It is based on contrastive learning and is efficient in applications such as cross-modal retrieval, image classification, and visual question answering [1].

This work presents a comparative study on fine-tuning BioMedCLIP for specialized medical tasks. Its efficiency was evaluated using different techniques and various developments across two challenging datasets, namely, the National Institutes of Health (NIH) Chest X-ray collection for multi-label disease classification and the HAM10000 dataset for skin lesion analysis. Through experimentation with various fine-tuning strategies, the aim was to enhance the model's capacity to integrate imaging and textual data while improving its performance on clinical tasks. The findings validate the effectiveness of domain-specific adaptation and highlight the potential of BioMedCLIP as a robust tool for advancing AI-driven healthcare applications [1].

II. RELATED WORK

The coordination between various modalities in AI is driving a paradigm shift in multiple research sectors, especially in biomedical research and clinical practice. These developments are undoubtedly contributing to the digitalization and virtualization of healthcare, promising to revolutionize personalized healthcare on a global scale.

The concept of a virtual hospital was clearly proposed in [2], which discussed integrating discriminative AI and multimodal generative AI. It envisioned AI-powered continuous monitoring and precision intervention. However, it provided only a conceptual framework, lacking implementation, which remains a challenge along with interoperability, patient data privacy, and regulatory validation. Another study [3] aimed at mapping technical trends in multimodal machine learning for healthcare. This survey identified key architectural patterns in multimodal biomedical AI, concluding with a description of various challenges like data heterogeneity, missing modalities, and explainability. The highlighted gaps include lack of large cross-modal biomedical benchmarks, limited solutions for missing or noisy modalities, and the need for transparent techniques. Another study [4] proposed a contrastive attention mechanism for multimodal clinical prediction and was designed to handle missing modalities. However, it faced challenges such as scalable training and small dataset sizes. Two more studies, [5] and [6], conducted topic modeling and revealed research hotspots in multimodal AI and related systems, respectively. Additionally, an earlier study [7] performed pneumonia detection in chest X-rays using transfer learning.

Having established the context, we now review several relevant models. Contrastive Language–Image Pretraining (CLIP) [8], introduced by Open AI, is a multimodal neural network that trains an image encoder and a text encoder to learn visual concepts. It utilized a contrastive learning framework, aligning images with their corresponding text captions in a shared embedding space by maximizing similarity of matched pairs and minimizing incorrect pairs. It was pretrained on a large-scale dataset of 400 million (image, text) pairs collected from the internet, enabling zero-shot learning on a variety of downstream image classification tasks without fine-tuning. MedCLIP [9] is a contrastive learning model trained on unpaired medical images and free-text reports. It leveraged large-scale unpaired data from PubMed and medical image datasets (MIMIC, CheXpert) and incorporated medical knowledge priors into the representation space. It was also evaluated on multiple medical imaging tasks, showing superior pathology classification compared to text-only or vision-only baselines. However, it relied on limited modality diversity, showed reduced performance for complex reasoning, performed less effectively on benchmarks, and had limited exploration of multilingual clinical text. PubMedCLIP [10] fine-tuned Open AI CLIP on the ROCO dataset and achieved

strong performance on radiology-specific tasks. However, its generalization was weaker than BioMedCLIP, and the limited dataset hindered multimodal performance. IBM Biomedical Foundation models [11] are general-purpose multi-omics and molecular biomedical foundation models, showing applications in drug discovery, molecular interaction predictions, and oncology target identification. However, they demonstrated less integration with Electronic Health Records (EHR) and imaging data and were largely closed source.

This paper focuses on BioMedCLIP, a powerful model by Microsoft [1]. It is a state-of-the-art model, and this study presents a comparative analysis that adapts BioMedCLIP and evaluates its efficiency across various datasets and fine-tuning approaches. We worked on fine-tuning techniques and evaluated the model on two large and challenging datasets.

III. DATASET DESCRIPTION

To study the developments performed on BioMedCLIP comparatively, two datasets were employed. Firstly, the NIH Chest X-ray dataset is a large-scale public collection of 112,120 frontal-view X-ray images from 30,805 unique patients [12]. The dataset provides weakly supervised labels for 14 common thoracic pathologies, including Atelectasis, Cardiomegaly, Effusion, Infiltration, Mass, Nodule, Pneumonia, Pneumothorax, Consolidation, Edema, Emphysema, Fibrosis, Pleural Thickening, and Hernia. These labels were generated automatically by applying Natural Language Processing (NLP) to the associated radiology reports, with an estimated accuracy of over 90%. The dataset is organized with patient-level predefined splits for training and testing to ensure robust and fair model evaluation. Figure 1 shows sample images from the NIH Chest X-ray dataset.

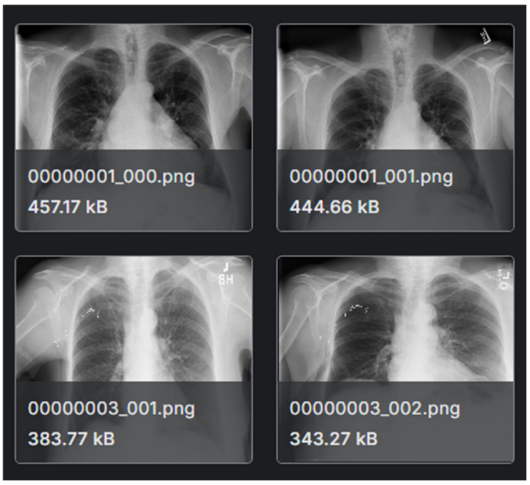

Fig. 1. Sample images from the NIH Chest X-ray dataset.

Secondly, the HAM10000 dataset [13] is a large public collection of 10,015 dermatoscopic images. It provides ground-truth labels for seven common diagnostic categories of pigmented skin lesions. These include Actinic keratoses (akiec), basal cell carcinoma (bcc), benign keratosis-like lesions (bkl), dermatofibroma (df), melanoma (mel), melanocytic nevi (nv), and vascular lesions (vasc). The diagnostic labels are highly reliable, as over 50% of cases were confirmed by histopathology, with the remainder verified by expert consensus or follow-up examinations. The collection was curated to ensure diversity, featuring images from different populations acquired via various modalities. Figure 2 shows sample images from the HAM10000 dataset.

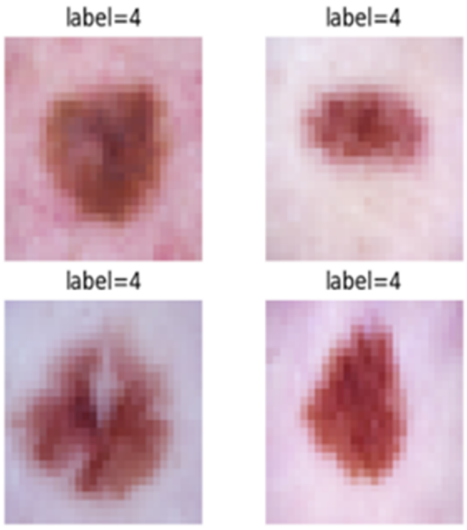

Fig. 2. Sample images from the HAM10000 dataset (skin lesions).

IV. METHODOLOGY

This section provides a detailed description of the systematic framework developed to investigate the adaptability and optimize the performance of the pretrained BioMedCLIP model across a diverse range of biomedical imaging datasets, enabling the model to serve as a base for a virtual hospital, which represents one of the future aspects of this study. The architecture of the BioMedCLIP model is illustrated in Figure 3. The research is structured as a comparative analysis of different transfer learning strategies to identify effective fine-tuning methods.

A. Comparative Baseline

To establish a comparative baseline, the intrinsic quality of pretrained features were tested for each downstream task. This was done through linear probing, in which first the backbone was frozen and only a newly appended classification head was trained. This approach measured the linear separability between the classes in the model's existing embedding space and provided a reference point for evaluating the efficiency of subsequent fine-tuning strategies.

B. Fine-Tuning

The first approach was end-to-end fine-tuning, in which all parameters of the pretrained backbone and classification head were updated jointly. The training loop had a forward pass for prediction and loss computation, and backward pass for gradient computation and loss minimization via backpropagation.

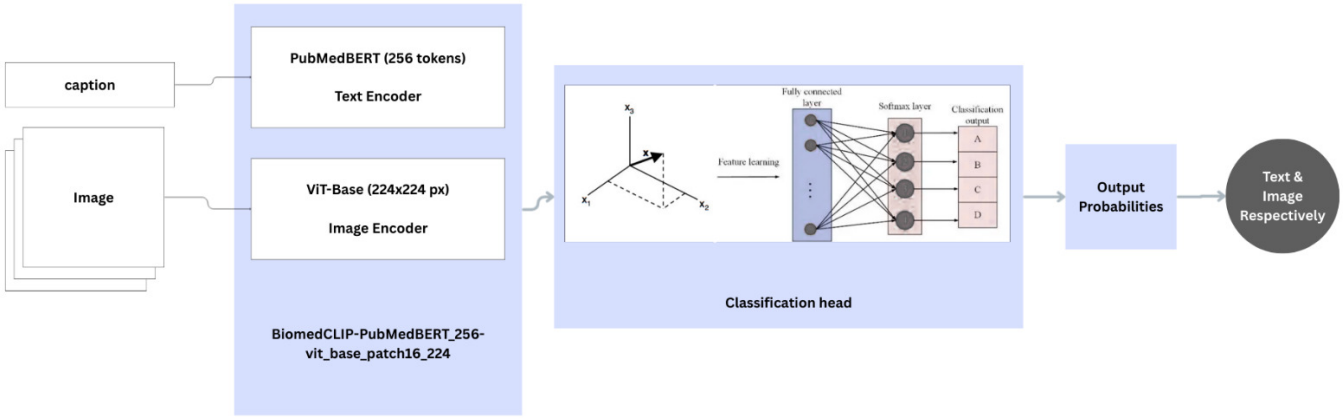

Fig. 3. Architecture of the BioMedCLIP model showing multimodal integration of image and text inputs.

The second approach was more sophisticated and focused on the multimodal fusion concept. It introduced a dedicated cross-modal fusion module positioned between the feature extractor and the final classifier. Unlike the previous approach, which relied solely on visual features, this module allowed the image and text embeddings to mutually refine one another. The image representation was contextualized by the text, and the text was grounded in the visual evidence from the image. The final, mutually informed embeddings were concatenated into a unified feature vector.

In the third approach, modifications were made in text prompting. Neural text prompts helped the model learn from the visual pathology rather than relying on textual shortcuts. The cross-attention fusion architecture was retained, but the output features from the image and text modalities were combined via averaging instead of concatenation before being passed to the final classifier. This creates a balanced, joint embedding for the classification task.

The model's parameters were partitioned into two groups. The pretrained BioMedCLIP encoders were trained with a very low learning rate (1e-5), whereas the newly initialized fusion and classification heads were trained with a significantly higher learning rate (1e-3). To address the low prevalence of certain diseases in the dataset, a weighted Binary Cross-Entropy (BCE) loss was employed. The loss for each class was weighted by the inverse frequency of its positive samples in the training set, giving greater importance to correctly classifying rarer conditions. To improve computational efficiency and reduce memory usage, Automatic Mixed Precision (AMP) was utilized during the training loop.

C. Equations

The proposed framework in this study can be represented mathematically as follows:

- Normalized feature extraction:

$$E_{img}(I) = \frac{f_{img}(I)}{\|f_{img}(I)\|} \quad (1)$$

$$E_{txt}(T) = \frac{f_{txt}(T)}{\|f_{txt}(T)\|} \quad (2)$$

where $I \in \mathbb{R}^{H\times W\times 3}$ and $T \in \mathbb{R}^{d}$.

- Cross–modal fusion:

$$F = F_{fusion}\left(E_{img}(I), E_{txt}(T)\right) \quad (3)$$

- Classification:

$$P = \sigma\big(C(F)\big) \quad (4)$$

V. EXPERIMENTAL RESULTS

This section presents the observations and evaluation results obtained from the procedures described above. To determine the relevance for biomedical applications, additional metrics were calculated for the last regularized multimodal approach. The model achieved an F1-score (macro) of 0.1772, F1-score (micro) of 0.1967, AUC (macro) of 0.6799, AUC (micro) of 0.7016, sensitivity (macro) of 0.6791, and sensitivity (micro) of 0.7546. These outcomes indicate moderate variability and suggest that further improvements are needed for the model across both datasets, reinforcing the potential biomedical relevance.

Table I presents the performance of the base model on different datasets to provide a reference for comparison with subsequent fine-tuning approaches.

TABLE I. BASELINE MODEL PERFORMANCE

Dataset	Description	Test accuracy (%)
Chest X-ray	BioMedCLIP encoders were used for baseline training without fine-tuning. The model's accuracy was low, as it struggled to learn meaningful representations	25
HAM10000	Default training on the HAM10000 dataset provided moderate accuracy without optimization	60

Table II presents the results of the first approach, which involved end-to-end fine-tuning in which all model weights were updated based on performance. Table III presents the accuracy results for the second approach, which employed cross-attention fusion to integrate image and text embeddings. Finally, Table IV presents the accuracy results for the third approach, which employed regularized multimodal fine-tuning,

incorporating neural text prompts and differential learning rates to reduce overfitting and information leakage.

TABLE II. END-TO-END FINE-TUNING RESULTS

Dataset	Description	Test accuracy (%)
Chest X-ray	Model performed well during training but suffered severe overfitting in validation	55
HAM10000	Fine-tuning increased accuracy; nonetheless, a small degree of overfitting was observed	57

TABLE III. CROSS-ATTENTION FUSION RESULTS

Dataset	Description	Test accuracy (%)
Chest X-ray	Training and validation accuracy reached 99%, but testing results still indicated overfitting	44.9
HAM10000	Accuracy improved, nonetheless a small amount of overfitting persisted	66.9

TABLE IV. REGULARIZED MULTIMODAL RESULTS

Dataset	Description	Test Accuracy (%)
Chest X-ray	Improvement over previous approaches; reduced overfitting through neutral text prompts and differential learning rates	Micro acc: 55.6
		Macro acc: 55.6
HAM10000	Accuracy improved with regularization; neural prompts and differential learning rates mitigated overfitting	Micro acc: 83.67
		Macro acc: 66.51

A. Hyperparameter Justification and Evaluation Consistency

Various controlled experiments were conducted to select hyperparameters that balance training stability and generalization. Gradient explosion and overfitting were addressed by adjusting learning rates and weight decay. Differential learning rates were applied to fine-tune the encoder and classifier layers, and dropout was employed to improve regularization. After numerous revisions, these values were empirically chosen to provide consistent convergence and improved performance across both datasets. Table V presents the impact of these key hyperparameters on model performance and training efficiency.

To evaluate robustness, each experiment was repeated three times with different random seeds, and the results were averaged. Variation across runs was minimal (±1.5%), reinforcing the stability of the proposed fine-tuning strategies. Extensive cross-run testing will be conducted in future validation studies, as current GPU constraints limited this analysis.

VI. CHALLENGES AND LIMITATIONS

Several challenges and limitations were encountered in this study. Medical imaging datasets are often diverse and complex, making it difficult to organize data efficiently for model training. Scalability remains a concern, as applying the proposed method to larger datasets or additional modalities requires substantial modifications to preprocessing and model architecture. Computational resources also imposed restrictions, limiting the selection of model configurations, the number of training iterations, and batch sizes due to limited GPU memory and processing capacity. Lastly, the underrepresentation of certain pathologies in the datasets affected the model's ability to generalize across all classes, resulting in biased learning outcomes.

TABLE V. HYPERPARAMETER COMPARISON

Experiment	Parameter modified	Original value(s)	New value(s)	Accuracy (%)	Training time (s)
E1 (baseline)	Default model	Optimizer	No change	25	20.22
E2	Retrained default model	Zero grad	No change	55	18.77
E3	Learning rate, seed, dropout	2e-4, 42, 0.1	1e-2, 42, 0.5	44.9	17.33
E4	WEIGHT_DECAY, LR_HEAD	1e-2, 1e-3	1e-4, 1e-2	Micro: 55.6; Macro: 55.6	16.89
E5	LR_ENCODERS, LR_HEAD	1e-5, 1e-3	1e-2, 1e-2	Micro: 83.67; Macro: 66.51	13.87

VII. FUTURE DIRECTIONS

The study demonstrates successful fine-tuning and evaluation of BioMedCLIP, laying the groundwork for several promising avenues for future research. However, the long-term vision extends beyond static image analysis toward integrated clinical support, forming a foundation for virtual hospital concepts. A significant next step is incorporating sequential learning, as the current models are designed for static datasets. This would enable the model to move beyond single diagnostic snapshots toward more dynamic, continuous problem-solving. As discussed, the ultimate goal is to validate BioMedCLIP as a core component within a larger integrated platform. Although the proposed approaches yielded strong internal results, there was a trend of slight overfitting. Therefore, external validation on datasets such as CheXpert, MIMIC-CXR, and SIC is planned to confirm generalization, which was not performed in this study due to GPU resource constraints. Additional future aspects include applying the model to large-scale datasets with thousands of classes, system-level integration, designing an intuitive user interface, and conducting rigorous real-world trials to assess the impact on clinical decision-making.

VIII. CONCLUSION

This research successfully investigated the foundational model BioMedCLIP and its adaptation to diverse biomedical datasets. The primary objective was to fine-tune the model using different strategies and evaluate their effects on performance. Experiments were conducted on two distinct datasets: the National Institutes of Health (NIH) Chest X-ray collection and the HAM10000 dermatology dataset. The results validate that strategic fine-tuning can transform a general-purpose biomedical model into a more specialized tool, capable

of supporting effective clinical decision-making. The comparative study indicated that, although full fine-tuning improved performance, it also resulted in overfitting. Validation accuracy improved, but stability remained problematic in the case of cross-attention fusion. The regularized multimodal approach, which combined neural text prompts with differential learning rates, achieved the best-balanced performance, mitigating the limitations observed in other methods.

For biomedical applications of BioMedCLIP, controlled fine-tuning proved the most efficient. Specifically, three approaches were tested: the first focused on end-to-end simple fine-tuning, the second experimented with parameter-efficient fine-tuning, and the third applied a more structured hyperparameter tuning for both the classifier head and the backbone. This study establishes a critical concept for future adaptation of BioMedCLIP in advanced clinical systems, serving as a foundation for advanced multimodal digital healthcare and the development of virtual hospital platforms.

REFERENCES

[1] S. Zhang *et al.*, "BiomedCLIP: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs." arXiv, Jan. 08, 2025, https://doi.org/10.48550/arXiv.2303.00915.

[2] M. A. Rahman and S. Al-Hazzaa, "Next-Generation Virtual Hospital: Integrating Discriminative and Large Multi-Modal Generative AI for Personalized Healthcare," in *GLOBECOM 2024 - 2024 IEEE Global Communications Conference*, Cape Town, South Africa, 2024, pp. 3509–3514, https://doi.org/10.1109/GLOBECOM52923.2024.10901624.

[3] S. Agrawal *et al.*, "Selection of 51 predictors from 13,782 candidate multimodal features using machine learning improves coronary artery disease prediction," *Patterns*, vol. 2, no. 12, Dec. 2021, Art. no. 100364, https://doi.org/10.1016/j.patter.2021.100364.

[4] J. Liu, D. Capurro, A. Nguyen, and K. Verspoor, "Attention-based multimodal fusion with contrast for robust clinical prediction in the face of missing modalities," *Journal of Biomedical Informatics*, vol. 145, Sept. 2023, Art. no. 104466, https://doi.org/10.1016/j.jbi.2023.104466.

[5] X. Chen, H. Xie, X. Tao, F. L. Wang, M. Leng, and B. Lei, "Artificial intelligence and multimodal data fusion for smart healthcare: topic modeling and bibliometrics," *Artificial Intelligence Review*, vol. 57, no. 4, Mar. 2024, Art. no. 91, https://doi.org/10.1007/s10462-024-10712-7.

[6] Q. Cai, H. Wang, Z. Li, and X. Liu, "A Survey on Multimodal Data-Driven Smart Healthcare Systems: Approaches and Applications," *IEEE Access*, vol. 7, pp. 133583–133599, 2019, https://doi.org/10.1109/ACCESS.2019.2941419.

[7] N. C. Kundur, B. C. Anil, P. M. Dhulavvagol, R. Ganiger, and B. Ramadoss, "Pneumonia Detection in Chest X-Rays using Transfer Learning and TPUs," *Engineering, Technology & Applied Science Research*, vol. 13, no. 5, pp. 11878–11883, Oct. 2023, https://doi.org/10.48084/etasr.6335.

[8] A. Radford *et al.*, "Learning Transferable Visual Models From Natural Language Supervision," in *Proceedings of the 38th International Conference on Machine Learning*, Online, 2021, pp. 8748–8763.

[9] Z. Wang, Z. Wu, D. Agarwal, and J. Sun, "MedCLIP: Contrastive Learning from Unpaired Medical Images and Text," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Abu Dhabi, United Arab Emirates, 2022, pp. 3876–3887, https://doi.org/10.18653/v1/2022.emnlp-main.256.

[10] S. Eslami, C. Meinel, and G. de Melo, "PubMedCLIP: How Much Does CLIP Benefit Visual Question Answering in the Medical Domain?," in *Findings of the Association for Computational Linguistics: EACL 2023*, Dubrovnik, Croatia, 2023, pp. 1181–1193, https://doi.org/10.18653/v1/2023.findings-eacl.88.

[11] "Biomedical Foundation Models." IBM Research. https://research.ibm.com/projects/biomedical-foundation-models?utm_source=chatgpt.com.

[12] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases," in *2017 IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, 2017, pp. 3462–3471, https://doi.org/10.1109/CVPR.2017.369.

[13] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Scientific Data*, vol. 5, no. 1, Aug. 2018, Art. no. 180161, https://doi.org/10.1038/sdata.2018.161.