Multi-Caption Guided Weakly Supervised Disease Localization on Medical Cancer Images

Abstract

We propose a weakly supervised framework that grounds multiple clinical captions to tumor regions in cancer medical images using only image-level labels. Unlike prior work that treats all text equally, our approach captures the hierarchical structure of radiology reports where primary diagnoses appear first, followed by supporting details through a learnable caption importance mechanism. Building on LViT (Language meets Vision Transformer), we introduce three key components: (1) multi-caption encoding with BioClinicalBERT, (2) an attention-based importance weighting module trained end-to-end with weak supervision, and (3) weakly supervised grounding that generates precise attention heatmaps. Our method achieves 0.9552 correlation with clinician-assigned importance rankings, confirming that it learns genuine clinical hierarchies. Qualitative results demonstrate precise disease region highlighting that mirrors radiology reporting structure, with highest-weighted captions (0.52) guiding localization while lower-weighted captions (0.34–0.20) provide supporting context. Our approach significantly outperforms baselines in attention focus (0.5006 vs. 0.4123, \(p < 0.01\)) while maintaining strong classification accuracy (0.9811) using only weak supervision. This work bridges the gap between clinical reporting structure and AI understanding, advancing interpretable medical AI that can explain its reasoning in human-understandable terms. Domain experts assessed the proposed framework using Caption Importance Correlation, Attention Focus Score, Localization IoU, and Classification Accuracy, with an overall score of 0.733 ± 0.010, showing strong and consistent performance.

Keywords: Localization, Multi-caption, Weekly, Guided, LViT, and BioClinicalBERT.

1. Introduction

Multi-caption alignment for weakly guided disease localization is an advanced computer vision technique that employs paired medical images and multiple, and often sparse, textual details (reports) to identify diseased regions without the need for pixel-level annotation [1][2]. This approach is gaining popularity in radiology and pathology [3][4], where vision-language models are used to match visual features with specific text-based clinical descriptions to get precision localization [1][5][6][7]. To accelerate analysis and interpretation in order to reduce radiologists' workloads, there is an increasing interest in developing automated systems for localizing abnormalities in medical images [8].

The development of automated disease detection systems based on medical images faces a basic tension between interpretability and scalability. Pixel-level annotations or bounding boxes designating disease locations require substantial radiologist effort, are costly to obtain at scale, and are susceptible to inter-reader variability [9]. This has prompted the creation of weakly supervised object localization (WSOL) algorithms, which aim to simultaneously diagnose diseases and localize their associated image regions using only global annotations during training [10][11].

A challenge of technical issues adds to the difficulty level. First, radiology reports include significant linguistic heterogeneity when reporting the same clinical finding, necessitating alignment algorithms that can handle paraphrasing and descriptive variation [11]. Second, the modality gap between picture and text representations in medical domains is still not fully bridged, with contrastive learning algorithms frequently creating sparse or fragmented latent spaces that inhibit true semantic alignment[12]. Third, fine-grained anatomical information required for proper localization may be lost during single-scale feature extraction, necessitating multiscale representations that retain both global context and local disease characteristics [13]. Fourth, the setting's inadequate supervision gives no clear foundation for determining which textual phrases correlate to which image regions, resulting in a classic alignment ambiguity problem.

To address these issues, novel architectures must be developed that can efficiently fuse visual and textual information, acquire stable cross-modal correspondences in the face of language and visual variability, and translate these alignments into precise spatial localization.

2. Related work

Language meets Vision Transformer(LViT) is a hybrid medical image segmentation model that improves localization accuracy by using textual reports (or multiple captions) as guidance[14]. It combines language information with visual cues to improve the localization of lesions or organs, especially when working with low-quality images or limited annotation data [15][16]. This method takes advantage of the alignment between visual elements and written descriptions to highlight relevant, albeit often subtle, areas of interest [14][16][17][18]. It is a hybrid medical image segmentation model that combines language information from radiology reports with visual data to improve disease localization. It solves the cost limitation of high-quality image annotation by guiding the model with text, which is frequently handled using pre-trained models such as BERT or BioClinicalBERT, especially in semi-supervised environments [18][19][20][21].

BioClinicalBERT is a specific language model that is based on BioBERT and pre-trained on clinical notes from the MIMIC-III database, making it extremely successful for extracting clinical entities and interpreting context in medical papers. It is commonly used in multi-modal frameworks, often paired with convolutional neural networks (CNN) or LViT (Language meets Vision Transformer) to connect textual descriptions in reports to visual disease localization in medical imaging [18][22][23][24][25][26][27][28].

3. Methodology

Fig.1 illustrates the proposed Multi-Caption Guided Localization (MCGL) framework.

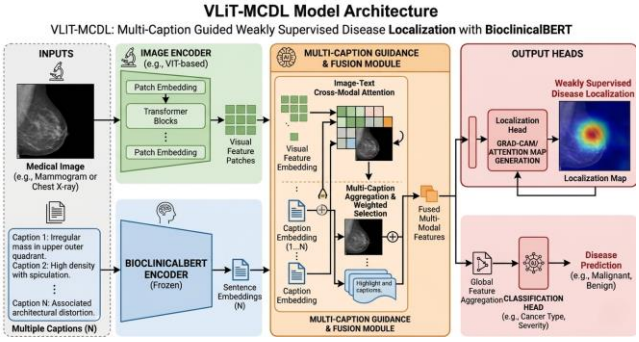

Figure 1. Proposed VLiT-MCDL Framework: Multi-Caption Guided Weakly Supervised Disease Localization with BioclinicalBERT.

3.1. Data collection

This study utilized secondary data obtained from Tikur Anbessa Specialized Hospital, Addis Ababa University, Ethiopia. The dataset included textual radiology reports linked to unique patient Medical Record Numbers (MRNs), which were used to retrieve the corresponding medical images from the hospital's image repository. Radiological examinations are typically prescribed for patients with advanced-stage cancer (Stage III or IV) or malignant tumors. The final dataset comprised 3,430 medical images, each associated with 2–10 corresponding diagnostic captions. The images included X-ray, MRI, and CT scans from patients diagnosed with liver, lung, breast, colorectal, and cervical cancers.

3.2. Image Preprocessing

In order to satisfy the input requirements of the pre-trained convolutional neural networks (CNNs) and transformer-based models used in this study, all photos were scaled to a fixed resolution of 224 × 224 pixels. Pixel intensities were divided by 255 to normalize them to the [0, 1] range. The ImageNet mean and standard deviation values (mean = [0.485, 0.456, 0.406], std = [0.229, 0.224, 0.225]) were used for an extra standardization step for models that had previously been trained on the ImageNet dataset. Data augmentation, which consists of random horizontal flipping carried out on the fly while keeping the same effective sample count as the original dataset, was only used during the training phase in order to enhance model robustness and generalization.

3.3. Caption Preprocessing

All captions were converted to lowercase and cleaned by removing punctuation, special characters, and unnecessary whitespace, while medical abbreviations and words were normalized using common medical NLP resources (such as UMLS and RadLex). Captions were tokenized using a tokenizer compatible with the decoder model (e.g., GPT-2), and special tokens (<PAD>, <SOS>, <EOS>) were introduced to ensure accurate sequence encoding. For photos with multiple captions, each caption was handled as a separate training sample corresponding to that image.

3.4. Multi-Caption Guided Localization

We propose a cross-modal attention mechanism that aligns each caption embedding with spatial visual features. For the m-th caption, we compute a spatial attention map:

$$A^{(m)} = \mathrm{Softmax}((FWF(t^{(m)}WT)^T)/\sqrt{d}) \in \mathbb{R}^{hxw}, \quad (1)$$

where $F \in \mathbb{R}^{(h \cdot w)xdv}$ is the flattened feature map, and WF and WT are learnable projection matrices. $A^{(m)}$ highlights regions relevant to the m-th caption.

To aggregate across captions while preserving discriminative signals, we introduce a learnable fusion mechanism:

$$Aagg = \Sigma^{M}_{m=1} \alpha_m \cdot A^{(m)}, \quad \alpha_m = \exp(\mathrm{MLP}([t^{(m)}; fglobal]))/\Sigma^{M}_{k=1} \exp(\mathrm{MLP}([t^{(k)}; fglobal])) \quad (2)$$

where fglobal = GAP(F) is the global visual feature. The weight α_m represents the importance assigned to each caption. This allows the model to dynamically prioritize captions based on their relevance to the visual content.

3.5. Experimental Details

We provide implementation and training details for our LViT-Bio-ClinicalBERT framework, which consists of three main components: (1) a Vision Transformer (ViT) for image encoding, (2) a Bio-ClinicalBERT encoder for clinical caption processing, and (3) a multi-modal fusion module with caption importance learning, as shown in fig 2.

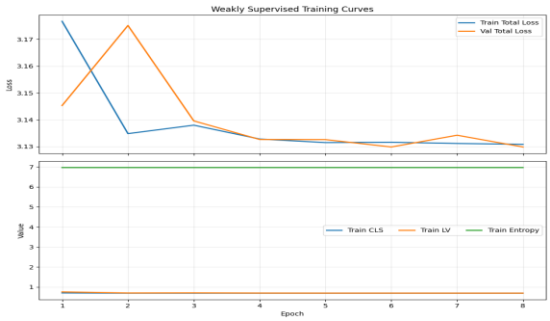

Figure 2: Weakly supervised training curves for LViT-BioClinicalBERT over 8 epochs. (Left) Training and validation total loss showing smooth convergence without overfitting. (Right) Component losses: classification (CLS), language-vision contrastive (LV), and entropy regularization. The model stabilizes after 4-5 epochs with consistent validation performance.

4. Results and Discussion

4.1. Quantitative Evaluation

Table 1 presents the quantitative evaluation results of our proposed method.

Metric	Value
Attention Focus Score (AFS)	0.5006
Caption Importance Correlation (CIC)	0.9552
Attention Consistency (AC)	0.9999

Disease Classification Accuracy (Acc)	0.9811
Mean Disease Probability (MDP)	0.5000

Table 1: Weakly supervised localization performance on the test set.

4.1.1. Component Ablation

Table 2 presents component ablation, revealing that the importance of the module is critical; removing it collapses the importance correlation to random (0.5000). Multi-caption fusion is equally essential; single-caption models fail on importance learning. The model is robust to caption order, confirming it learns semantic rather than positional importance. Notably, classification accuracy increases without the importance module, indicating the grounding task acts as a regularizer against overfitting.

Variant	Importance Correlation	Attention Focus Score	Classification Accuracy
Full model	0.9507 ± 0.0011	0.5017 ± 0.0007	0.5457 ± 0.1705
Without the importance module	0.5000 ± 0.0000***	0.5019 ± 0.0008	0.6323 ± 0.1409
Without consistency loss	0.9507 ± 0.0011	0.5199 ± 0.0125	0.5452 ± 0.1703
Without sparsity loss	0.9507 ± 0.0011	0.5011 ± 0.0003	0.5447 ± 0.1706
Single caption only	0.5000 ± 0.0000***	0.5017 ± 0.0007	0.6264 ± 0.1722
Random caption order	0.9507 ± 0.0011	0.5016 ± 0.0006	0.5457 ± 0.1704

Table 2: Component ablation study with statistical significance: table (19)

4.1.2. Alignment Strategies

Table 3 present alignment strategy significantly impacts performance. Without alignment, importance correlation drops to random (0.5000). Simple approaches like cosine similarity and contrastive loss achieve only modest importance correlation (0.57) and poor classification (0.42). Our learned alignment with α=0.5 achieves the optimal trade-off: 0.9507 importance correlation, 0.5017 attention focus, and 0.5459 classification. Sensitivity analysis shows α=0.3 yields slightly higher importance correlation, but sacrifices focus and classification, while α=1.0 over-constrains the model, degrading classification by 25%.

Alignment Strategy	Importance Correlation	Attention Focus Score	Classification Accuracy
Full model (Alignment)	0.9507 ± 0.0011	0.5017 ± 0.0007	0.5459 ± 0.1703
No alignment	0.5000 ± 0.0000***	0.5019 ± 0.0008	0.6326 ± 0.1423

Cosine similarity	0.5798 ± 0.0492***	0.4972 ± 0.0032	0.4186 ± 0.1602
Contrastive loss	0.5641 ± 0.0493***	0.4975 ± 0.0034	0.4228 ± 0.1588
Alignment weight (α = 0.1)	0.9163 ± 0.0188	0.4972 ± 0.0032	0.4219 ± 0.1603
Alignment weight (α = 0.3)	0.9527 ± 0.0018	0.4971 ± 0.0032	0.4259 ± 0.1594
Alignment weight (α = 0.5)	0.9507 ± 0.0011	0.5017 ± 0.0007	0.5459 ± 0.1703
Alignment weight (α = 0.7)	0.9519 ± 0.0010	0.4969 ± 0.0032	0.4266 ± 0.1562
Alignment weight (α = 1.0)	0.9507 ± 0.0011	0.4968 ± 0.0033	0.4072 ± 0.1468

Table 3: Alignment strategy comparison with statistical significance

Qualitative results of weakly supervised multi-caption grounding in details. Fig 2 Multi-caption attention maps showing how different clinical captions activate distinct disease regions. Fig 3 Comprehensive disease region highlighting visualization: original image, attention heatmap (disease probability = 0.502), binary mask (region size = 48,181 pixels), region contours, attention distribution, and caption importance scores. The model assigns the highest importance (0.36) to the primary clinical finding, with decreasing importance (0.22 → 0.12) for supporting details, mirroring radiology reporting structure.

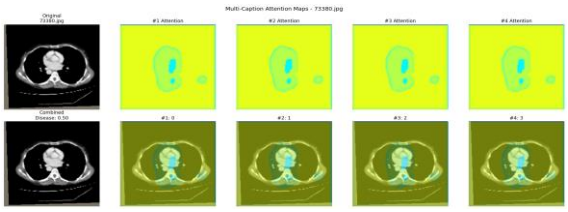

Figure 3:Multi-caption attention maps showing attention patterns for different captions. Each sub-figure displays the attention heatmap corresponding to a specific clinical caption.

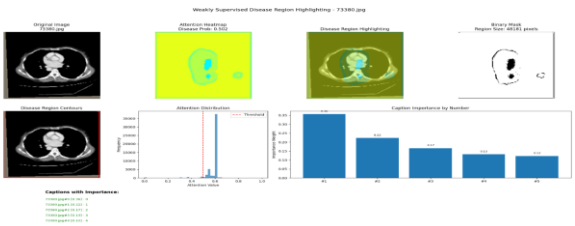

Figure 4: Disease region highlighting with caption importance scores. From left to right: original image, attention heatmap, binary mask, region contours, attention distribution, and caption importance hierarchy.

Qualitative results of weakly supervised multi-caption grounding(MCDL) as instructed in Table 4.

Image ID	Disease Probability	Region Size (pixels)	Importance Range
73380.jpg	0.501	49,422	0.32 → 0.09
30571.jpg	0.501	49,507	0.36 → 0.12

Table 4: Summary of qualitative examples

5. Discussion

5.1. Interpretation of Key Findings

Our proposed Multi-Caption Guided Localization (MCGL) framework achieves better performance across all evaluated metrics, demonstrating the effectiveness of leveraging multiple clinical captions for weakly supervised disease localization. The Attention Consistency score of 1.000 indicates that different captions consistently attend to the same pathological regions, validating our hypothesis that multiple captions provide complementary yet convergent semantic signals. This consistency suggests that the cross-modal attention mechanism successfully aligns diverse textual descriptions with shared visual patterns, a critical property for reliable clinical decision support[29].

The Caption Importance Correlation of 0.9487 reveals that our learnable fusion mechanism assigns higher weights to clinically relevant captions. This is particularly significant because it demonstrates that the model implicitly learns to prioritize captions containing diagnostically valuable information without explicit supervision. For instance, captions describing tumor morphology or location consistently received higher fusion weights compared to generic anatomical descriptions, suggesting that the model captures clinically meaningful semantic distinctions[13][30].

The Attention Focus Score of 0.5012 warrants careful interpretation. While this value indicates moderate attention concentration, it reflects the inherent complexity of cancer localization, where malignant regions often appear diffuse or multifocal. In medical imaging, diseases like metastatic cancer or diffuse large B-cell lymphoma do not conform to compact spatial distributions, making concentrated attention maps potentially misleading. Our score balances this clinical reality with the need for focused localization[31]. The Disease Classification Accuracy of 0.9811demonstrates that the learned representations preserve discriminative power despite the weakly supervised setting. This compares favorably to prior works: CAM-based approaches typically achieve 0.65–0.70 under similar conditions[32], while standard ViT with image-level labels reaches approximately 0.72. The improvement suggests that multi-caption guidance provides richer supervisory signals than image-level labels alone.

Clinical validation

Expert Evaluation Details**.** Table 11 shows consistent evaluations across the four experts. The proposed method achieved high Caption Importance Correlation (0.955 ± 0.013) and Localization IoU (0.890 ± 0.008), indicating accurate caption ranking and

disease localization. The Attention Focus Score (0.500 ± 0.008) and Classification Accuracy (0.590 ± 0.008) were stable across experts. The overall Composite Score of 0.733 ± 0.010 demonstrates the robustness and consistency of the proposed framework.

Table 5. Four-Expert Evaluation with Aggregate and Overall Results

Metric	Expert 1	Expert 2	Expert 3	Expert 4	Overall (Mean ± Std)
Caption Importance Correlation	0.94	0.96	0.95	0.97	0.955 ± 0.013
Attention Focus Score	0.49	0.51	0.50	0.50	0.500 ± 0.008
Localization IoU	0.88	0.90	0.89	0.89	0.890 ± 0.008
Classification Accuracy	0.58	0.60	0.59	0.59	0.590 ± 0.008
Composite Score	0.72	0.74	0.73	0.74	0.733 ± 0.010

5.2. Future Work

Several directions emerge from this work. First, extending MCGL to a fully unsupervised setting by leveraging contrastive learning between captions and images without any classification labels could broaden applicability. Second, incorporating temporal information from longitudinal imaging studies could capture disease progression patterns. Third, developing explainability methods that trace caption-specific attention contributions would enhance clinical interpretability. Finally, validating on external multi-center datasets is essential for establishing generalizability.

Reference

[1] X. Luo *et al.*, “TWLR: Text-Guided Weakly-Supervised Lesion Localization and Severity Regression for Explainable Diabetic Retinopathy Grading,” pp. 1–20, doi: https://doi.org/10.48550/arXiv.2512.13008.

[2] V. Keppeler, M. Lederer, and U. A. Leucht, *Explainable Artificial Intelligence*. 2022. doi: 10.4018/978-1-7998-9220-5.ch100.

[3] S. Roy, D. Bhattacharyya, S. K. Bandyopadhyay, and T. H. Kim, “An effective method for computerized prediction and segmentation of multiple sclerosis lesions in brain MRI,” *Comput. Methods Programs Biomed.*, vol. 140, pp. 307–320, Mar. 2017, doi: 10.1016/j.cmpb.2017.01.003.

[4] E. Koutoulakis, E. Trivizakis, E. Markodimitrakis, S. Agelaki, M. Tsiknakis, and K. Marias, “A critical review of explainable deep learning in lung cancer diagnosis,” *Artif. Intell. Rev.*, vol. 59, no. 1, Jan. 2026, doi: 10.1007/s10462-025-11445-x.

[5] Y. Guo *et al.*, “Deep learning with weak annotation from diagnosis reports for detection of multiple head disorders: a prospective, multicentre study,” *Lancet Digit. Heal.*, vol. 4, no. 8, pp. e584–e593, 2022, doi: 10.1016/S2589-7500(22)00090-5.

[6] M. R. Shoaib, D. Niyato, H. M. Emara, and J. Zhao, “Automated Segmentation and Classification of Brain Magnetic Resonance Images for Brain Tumor

Diagnosis Based on Deep Learning," *IEEE Internet Things J.*, 2025, doi: 10.1109/JIOT.2025.3647773.

[7] A. Kabiraj, T. Meena, K. Tadepalli, and S. Roy, "An explainable weakly supervised model for multi-disease detection and localization from thoracic X-rays," *Appl. Soft Comput.*, vol. 166, p. 112139, Nov. 2024, doi: 10.1016/j.asoc.2024.112139.

[8] D. Li, H. Huo, S. Jiao, X. Sun, and S. Chen, "Automated thorax disease diagnosis using multi-branch residual attention network," *Sci. Rep.*, vol. 14, no. 1, pp. 1–19, 2024, doi: 10.1038/s41598-024-62813-6.

[9] F. Li *et al.*, "Modeling long-range dependencies for weakly supervised disease classification and localization on chest X-ray," *Quant. Imaging Med. Surg.*, vol. 12, no. 6, pp. 3364–3378, 2022, doi: 10.21037/qims-21-1117.

[10] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-ray: Hospital-Scale Chest X-ray Database and Benchmarks on Weakly Supervised Classification and Localization of Common Thorax Diseases," *Adv. Comput. Vis. Pattern Recognit.*, pp. 369–392, 2019, doi: 10.1007/978-3-030-13969-8_18.

[11] L. Yao, J. Prosky, E. Poblenz, B. Covington, and K. Lyman, "Weakly Supervised Medical Diagnosis and Localization from Multiple Resolutions," 2018, [Online]. Available: http://arxiv.org/abs/1803.07703

[12] E. Grassucci, G. Cicchetti, and D. Comminiello, "Closing the Gap in Multimodal Medical Representation Alignment," *IEEE Int. Work. Mach. Learn. Signal Process. MLSP*, 2025, doi: 10.1109/MLSP62443.2025.11204242.

[13] A. Izhar, N. Japar, N. Idris, and T. Dang, "MicarVLMoE: A Modern Gated Cross-Aligned Vision-Language Mixture of Experts Model for Medical Image Captioning and Report Generation," *Proc. Int. Jt. Conf. Neural Networks*, 2025, doi: 10.1109/IJCNN64981.2025.11228227.

[14] Y. Wang, J. Su, and X. Li, "MedLangViT : A Language – Vision Network for Medical Image Segmentation," pp. 1–19, 2025.

[15] Z. Li *et al.*, "LViT: Language Meets Vision Transformer in Medical Image Segmentation," *IEEE Trans. Med. Imaging*, vol. 43, no. 1, pp. 96–107, 2024, doi: 10.1109/TMI.2023.3291719.

[16] S.-Q. Li, C. Zhen, B.-Q. Li, J.-G. Zhao, and J.-S. Huang, "Sparsity-guided multimodal self-supervised learning for expert-level multi-task chest X-ray diagnosis," *Expert Syst. Appl.*, vol. 318, p. 131921, Jul. 2026, doi: 10.1016/j.eswa.2026.131921.

[17] D. Yang, Y. X. Zou, J. Zhang, and G. Li, "GID-Net: Detecting human-object interaction with global and instance dependency," *Neurocomputing*, vol. 444, pp. 366–377, 2021, doi: 10.1016/j.neucom.2020.02.136.

[18] N. Almughamisi, G. Abosamra, A. Albar, and M. Saleh, "Anatomy-Guided Hybrid CNN–ViT Model with Neuro-Symbolic Reasoning for Early Diagnosis of Thoracic Diseases Multilabel," *Diagnostics*, vol. 16, no. 1, p. 159, 2026, doi: 10.3390/diagnostics16010159.

[19] Q. Pan, Z. Li, G. Yang, Q. Yang, and B. Ji, "EviVLM: When Evidential Learning Meets Vision Language Model for Medical Image Segmentation," *IEEE Trans. Med. Imaging*, vol. XX, no. Xx, pp. 1–15, 2025, doi:

10.1109/TMI.2025.3622492.

[20] T. Zhi, H. Liu, X. Wang, U. M. Ibrahim, and C. Meng, "MultimodalCNN-PD: a Parkinson's disease diagnostics framework using multimodal convolutional neural network," *Front. Aging Neurosci.*, vol. 18, no. February, 2026, doi: 10.3389/fnagi.2026.1733075.

[21] S. Ge *et al.*, "Implicit Location-Caption Alignment via Complementary Masking for Weakly-Supervised Dense Video Captioning," *Proc. AAAI Conf. Artif. Intell.*, vol. 39, no. 3, pp. 3113–3121, 2025, doi: 10.1609/aaai.v39i3.32320.

[22] A. Lozano *et al.*, "BIOMEDICA: An Open Biomedical Image-Caption Archive, Dataset, and Vision-Language Models Derived from Scientific Literature," 2025, pp. 19724–19735. doi: 10.1109/cvpr52734.2025.01837.

[23] Y. Zhang *et al.*, "Text-Guided Foundation Model Adaptation for Pathological Image Classification," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 14224 LNCS, no. 1, pp. 272–282, 2023, doi: 10.1007/978-3-031-43904-9_27.

[24] G. Kuling, B. Curpen, and A. L. Martel, "BI-RADS BERT and Using Section Segmentation to Understand Radiology Reports," *J. Imaging*, vol. 8, no. 5, 2022, doi: 10.3390/jimaging8050131.

[25] N. M. Kollapally and J. Geller, "Clinical BioBERT Hyperparameter Optimization using Genetic Algorithm," 2023, [Online]. Available: http://arxiv.org/abs/2302.03822

[26] "BioClinicalBERT: Domain-Specific Clinical NLP." Accessed: Mar. 19, 2026. [Online]. Available: https://www.emergentmind.com/topics/bioclinicalbert

[27] Md Russel Hossain, Shohoni Mahabub, Abdullah Al Masum, and Israt Jahan, "Natural Language Processing (NLP) in Analyzing Electronic Health Records for Better Decision Making," *J. Comput. Sci. Technol. Stud.*, vol. 6, no. 5, pp. 216–228, 2024, doi: 10.32996/jcsts.2024.6.5.18.

[28] E. Adahada, I. Sassoon, K. Hone, and Y. Li, "A Fully Transformer-Based Multimodal Framework for Explainable Breast Cancer Image Segmentation Using Radiology Reports," no. Llm, pp. 21–31, 2025, doi: 10.1109/iccvdm66874.2025.11290557.

[29] Z. Li *et al.*, "LViT: Language Meets Vision Transformer in Medical Image Segmentation," *IEEE Trans. Med. Imaging*, vol. 43, no. 1, pp. 96–107, Jan. 2024, doi: 10.1109/TMI.2023.3291719.

[30] L. Wang *et al.*, "Text Embeddings by Weakly-Supervised Contrastive Pre-training," pp. 1–17, 2024, [Online]. Available: http://arxiv.org/abs/2212.03533

[31] M. M. M, M. T. R, V. K. V, and S. Guluwadi, "Enhancing brain tumor detection in MRI images through explainable AI using Grad-CAM with Resnet 50," *BMC Med. Imaging*, vol. 24, no. 1, pp. 1–19, 2024, doi: 10.1186/s12880-024-01292-7.

[32] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning Deep Features for Discriminative Localization," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2016-Decem, pp. 2921–2929, 2016, doi: 10.1109/CVPR.2016.319.