MedConcept: Unsupervised Concept Discovery for Interpretability in Medical VLMs

Md Rakibul Haque[1,2,3], K M Arefeen Sultan[1,2], Tushar Kataria[1,2], and Shireen Y. Elhabian[1,2]

[1] Kahlert School of Computing, University of Utah, Salt Lake City, USA
[2] Scientific Computing and Imaging Institute, University of Utah, Salt Lake City, USA
[3] Corresponding author
{rakibul,arefeen.sultan,tushar.kataria,shireen}@sci.utah.edu

Abstract. While medical Vision-Language models (VLMs) achieve strong performance on tasks such as tumor or organ segmentation and diagnosis prediction, their opaque latent representations limit clinical trust and the ability to explain predictions. Interpretability of these multimodal representations are therefore essential for the trustworthy clinical deployment of pretrained medical VLMs.However, current interpretability methods, such as gradient- or attention-based visualizations, are often limited to specific tasks such as classification. Moreover, they do not provide concept-level explanations derived from shared pretrained representations that can be reused across downstream tasks. We introduce MedConcept, a framework that uncovers latent medical concepts in a fully unsupervised manner and grounds them in clinically verifiable textual semantics. MedConcept identifies sparse neuron-level concept activations from pretrained VLM representations and translates them into pseudo-report-style summaries, enabling physician-level inspection of internal model reasoning. To address the lack of quantitative evaluation in concept-based interpretability, we introduce a quantitative semantic verification protocol that leverages an independent pretrained medical LLM as a frozen external evaluator to assess concept alignment with radiology reports. We define three concept scores, Aligned, Unaligned, and Uncertain, to quantify semantic support, contradiction, or ambiguity relative to radiology reports and use them exclusively for post hoc evaluation. These scores provide a quantitative baseline for assessing interpretability in medical VLMs. Code is available at `https://github.com/RakibulHaqueSajal/MedConcept`.

Keywords: Unsupervised Concept Discovery · 3D Medical Image analysis · Report Generation

1 Introduction

Foundation models, particularly multimodal Vision-Language Models (VLMs) trained on large-scale medical datasets, are increasingly adopted across medical image analysis tasks such as tumor segmentation, disease classification,

and image-guided intervention [3,23,15]. Despite their strong performance, these models rely on complex latent representations that remain opaque to clinicians, limiting trust and interpretability [10,20]. This opacity is especially concerning in high-stakes settings such as radiology, surgical planning, and diagnostic decision making, where erroneous or misaligned representations may lead to significant clinical consequences [8,9]. Current interpretability techniques, including attention- and gradient-based visualizations, tend to focus on explaining individual task outputs while providing only partial visibility into the semantic organization of the underlying latent space [14,6,11]. As a result, there remains a pressing need for systematic methods that can recover clinically meaningful concepts directly from pretrained representations in a task-agnostic manner.

Sparse autoencoders (SAEs) have emerged as a powerful tool for unsupervised concept discovery in deep neural networks, first demonstrating their impact in large language models by decomposing representations into semantically meaningful and disentangled features [5,22,12]. Subsequent work has shown that SAEs similarly reveal sparse, interpretable structure in vision models, identifying object parts, textures, cellular structures, and clinically relevant patterns [19,2,13,7]. Building on this foundation, we introduce MedConcept, which applies sparse decomposition to pretrained medical VLM representations to uncover clinically meaningful latent factors without manual concept annotations. MedConcept grounds these latent representations in clinically verifiable textual medical semantics, linking neuron-level activations to semantically aligned medical concepts. The framework also produces patient-level concept summaries that clinicians can review to verify findings, refine annotations, or diagnose model errors, such as overlooked or inaccurately represented features. The entire concept discovery pipeline operates without task-specific supervision, paired image text training, or manual concept annotations. Public medical ontologies [17,4] are used to build the candidate concept vocabulary.

While concept discovery research has proposed various evaluation strategies, recent applications of SAEs in natural and medical imaging have relied primarily on manual inspection and selected visualizations rather than systematic quantitative assessment [19,2,13,7]. Such qualitative analyses are inherently limited in scale and susceptible to selection bias, making it difficult to determine whether discovered concepts capture clinically meaningful semantics or provide a reliable basis for benchmarking models. To address this gap in concept-based interpretability, we introduce a systematic LLM-based semantic entailment scoring framework for evaluating SAE-driven concept discovery in medical VLMs. Leveraging paired clinical reports and a frozen, independently pretrained medical LLM, the framework performs per-concept semantic inference to enable automated, scalable assessment of alignment between discovered concepts and textual evidence. We define three complementary scores, *Aligned, Unaligned*, and *Uncertain*, to quantify whether predicted concepts are supported by, contradicted by, or remain ambiguous with respect to the corresponding reference radiology reports. The reports are used exclusively for post hoc evaluation, while concept discovery and extraction remain fully unsupervised. This framework enables sys-

tematic benchmarking of interpretability in current models and provides a principled foundation for evaluating future medical vision–language models.

The main contributions of this paper are as follows:

- We introduce MedConcept, an unsupervised framework for medical concept discovery from pretrained VLM representations using sparse decomposition to identify clinically meaningful neuron-level latent factors.
- We ground discovered latent concepts in textual medical semantics by mapping sparse activations to aligned clinical concepts and generating patient-specific concept summaries.
- We propose a systematic LLM-based semantic entailment protocol to perform per-concept inference on paired clinical reports, introducing three complementary metrics, *Aligned, Unaligned, and Uncertain*, to quantify semantic support, contradiction, and ambiguity of discovered concepts

2 MedConcept

The proposed framework comprises three sequential stages: (1) *Concept discovery and extraction* via sparse autoencoder (SAE), (2) *Semantic grounding* of the discovered latent concepts through alignment with textual medical concepts dictionary (shown in Figure 1), and (3) *Quantitative evaluation* of concept–report semantic alignment.

Latent Concept Discovery and Extraction. Given a volumetric image $\mathbf{x} \in \mathbb{R}^{H\times W\times D}$, where H, W, and D denote the height, width, and depth (number of slices) of the volume, we extract a global feature embedding $\mathbf{v} = \Phi(\mathbf{x}) \in \mathbb{R}^p$

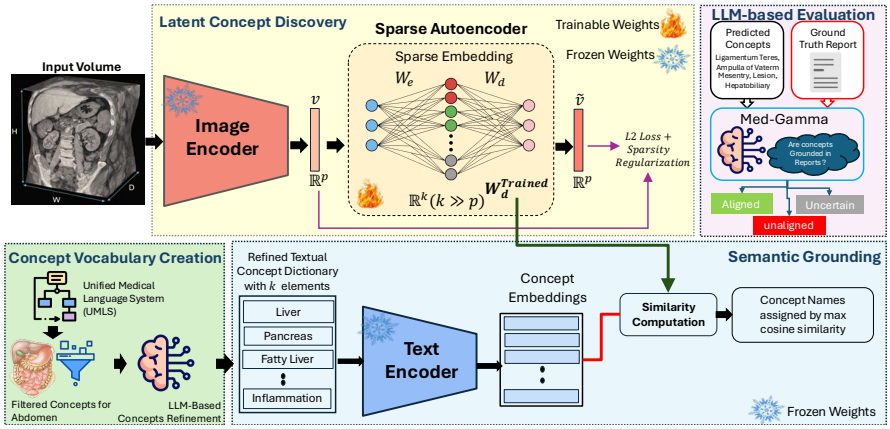

Fig. 1: MedConcept– **an unsupervised concept discovery and semantic verification framework.** A sparse autoencoder (SAE) produces sparse activations/embeddings that are matched with textual embeddings generated by a frozen text decoder over the full concept dictionary, assigning each neuron to the concept with the highest similarity.

from a pretrained 3D medical vision-language model (instantiated as Merlin [3] in our experiments), where p denotes the embedding dimensionality. Such volumetric embeddings are typically high-dimensional and poly-semantic, rendering individual dimensions neither semantically meaningful nor directly interpretable. To disentangle these representations into interpretable components, we train a sparse autoencoder (SAE) that maps the embedding into an over-complete sparse latent space. The encoder produces a sparse latent representation $\mathbf{z} \in \mathbb{R}^k$ and the decoder reconstructs the original embedding:

$$\mathbf{z} = \mathrm{ReLU}(\mathbf{W}_e\mathbf{v} + \mathbf{b}_e), \qquad \tilde{\mathbf{v}} = \mathbf{W}_d\mathbf{z} + \mathbf{b}_d. \tag{1}$$

Here, $\mathbf{W}_e \in \mathbb{R}^{k\times p}$ and $\mathbf{b}_e \in \mathbb{R}^k$ denote the encoder weights and bias, while $\mathbf{W}_d \in \mathbb{R}^{p\times k}$ and $\mathbf{b}_d \in \mathbb{R}^p$ denote the decoder weights and bias. The latent dimensionality k is chosen such that $k \gg p$, yielding an over-complete sparse representation. k is set equal to the cardinality of the candidate concept dictionary, ensuring that each neuron in the sparse autoencoder is matched to exactly one concept.

To encourage specialized and distinct latent factors, the SAE is trained using a unified objective comprising a reconstruction term and an ℓ_1 sparsity penalty on the activations:

$$\mathcal{L} = \underbrace{\|\mathbf{v} - \tilde{\mathbf{v}}\|_2^2}_{\text{Reconstruction}} + \lambda_1 \underbrace{\|\mathbf{z}\|_1}_{\text{Sparsity}}. \tag{2}$$

The sparsity term promotes activation selectivity, encouraging individual latent units to respond to specific structural patterns.

Textual Concept Vocabulary Creation. To create a dictionary of candidate concepts, we first extract relevant anatomical and pathological terms, along with their clinical synonyms, by querying the Unified Medical Language System (UMLS) using radiology-domain terminology. These terms are used exclusively to construct a candidate vocabulary for semantic grounding and are not used to supervise concept discovery or model training. Because these initial UMLS queries may not capture the full lexical diversity of clinical phrasing, we subsequently employ a large language model (ChatGPT) [18] to systematically expand the vocabulary with additional clinically relevant synonyms, thereby curating a comprehensive corpus of abdomen-specific concepts (Figure 1).

Concept Naming. An automated naming pipeline based on similarity matching is employed to align sparse autoencoder–discovered concepts with medical concepts in the candidate textual vocabulary dictionary, thereby assigning semantically interpretable labels to neuron activations. Let $\mathcal{T} = \{t_1, t_2, \ldots, t_N\}$ denote the curated vocabulary of N candidate terms. Each candidate term $t_i \in \mathcal{T}$ is projected into the joint vision–language latent space using the pretrained text encoder $\psi(\cdot)$ (instantiated as Merlin in our experiments), yielding a dense embedding $\mathbf{e}_i = \psi(t_i) \in \mathbb{R}^p$.

To interpret the j-th SAE concept unit, represented by its corresponding decoder dictionary vector $\mathbf{w}_j \in \mathbb{R}^p$ (the j-th column of $\mathbf{W}_d$), we compute the cosine similarity between the latent concept representation and all candidate

text embeddings:

$$\alpha_{ij} = \frac{\mathbf{w}_j^\top \mathbf{e}_i}{\|\mathbf{w}_j\|_2 \, \|\mathbf{e}_i\|_2}. \tag{3}$$

The semantic label $\hat{t}_j$ assigned to the j-th concept unit is determined by selecting the vocabulary term that maximizes this similarity:

$$\hat{t}_j = \arg\max_{t_i \in \mathcal{T}} \alpha_{ij}. \tag{4}$$

Quantitative Evaluation of Concept-Report Alignment. The proposed framework outputs textual concepts that are highly activated for a given patient image. However, qualitative comparisons by a medical profession increases cost and is prone to observer bias and limited in scalability. To quantitatively assess semantic alignment, we leverage MedGemma [21] as a frozen, independent evaluator to compare the predicted concept set $\mathcal{P}$ against the paired radiology report (note that reports are used for evaluation only).

For a given patient image $\mathbf{x}$, we define the predicted concept set $\mathcal{P}$ as the semantic labels of all active SAE neurons:

$$\mathcal{P} = \{\hat{t}_j \mid z_j > \tau, \ \forall j \in \{1, \cdots, k\}\},$$

where τ denotes a predefined activation threshold. If no concept units exceed the activation threshold, the image is excluded from evaluation.

For each predicted concept $p_i \in \mathcal{P}$, MedGemma performs per-concept semantic inference conditioned on the radiology report, producing a probability distribution over three clinical states:

$$\mathbf{s}_i = g_{\text{MedGemma}}(p_i, \mathcal{R}) \in \left[s_i^c\right]_{c \in \{\text{Aligned, Unaligned, Uncertain}\}}, \tag{5}$$

where $s_i^c \in [0, 1]$ and $\sum_c s_i^c = 1$. We obtain a discrete verdict $v_i = \arg\max_c s_i^c$ for each predicted concept and define normalized image-level alignment metrics:

$$\text{Score}(c) = \frac{1}{|\mathcal{P}|} \sum_{i=1}^{|\mathcal{P}|} \mathbf{1}(v_i = c), \quad c \in \{\text{Aligned, Unaligned, Uncertain}\}. \tag{6}$$

where $\mathbf{1}(\cdot)$ denotes the indicator function. By construction, the three scores sum to one, providing a normalized quantitative summary of semantic support, contradiction, and ambiguity relative to the radiology report.

3 Results

Datasets. We evaluate our framework on two large-scale 3D abdominal CT datasets to assess generalization across distributions and annotation types. AbdomenAtlas 3.0 [16] includes 9,262 volumes with radiology reports and per-voxel tumor annotations from 17 public datasets. Merlin Plus contains 25,494 volumes with dense anatomical annotations generated via R-Super [1].

For each dataset, we extract volumetric embeddings using the frozen Merlin encoder and train a separate sparse autoencoder (SAE) for dataset-specific concept discovery. SAE training uses only image embeddings, with radiology reports reserved for post hoc semantic evaluation.

Implementation Details. For each dataset, the SAE is trained on frozen feature representations extracted from the Merlin image encoder. All CT volumes are preprocessed using the same transformation pipeline as the pretrained encoder to ensure feature consistency. The SAE is optimized using Adam with a learning rate of 5×10^{-5}, an expansion factor of 2 (i.e., latent dimensionality $k = 2p$), and an ℓ_1 sparsity coefficient of 2×10^{-3}. Quantitative concept evaluation is performed using MedGemma-1.4B as a frozen semantic evaluator. GPT (ChatGPT v5.2) is used solely to expand and refine the UMLS-derived vocabulary during concept dictionary construction and is not involved in representation learning or evaluation.

Qualitative analysis. Figure 2 presents representative qualitative examples of MedConcept across best, median, and worst cases, ranked by the per-image Aligned score. Each case is visualized using axial, sagittal, and coronal views, alongside the top activated concepts grouped by semantic verification outcome (Aligned, Unaligned, Uncertain). In *high-scoring cases*, most predicted concepts are classified as Aligned, forming coherent anatomical and pathological clusters (e.g., hepatobiliary, vascular, mesenteric, ascites, congestion) that correspond to visually plausible structures in the CT volumes. In *median cases*, predictions are increasingly dominated by Unaligned concepts, often involving context-dependent or overly specific descriptors such as procedural details or

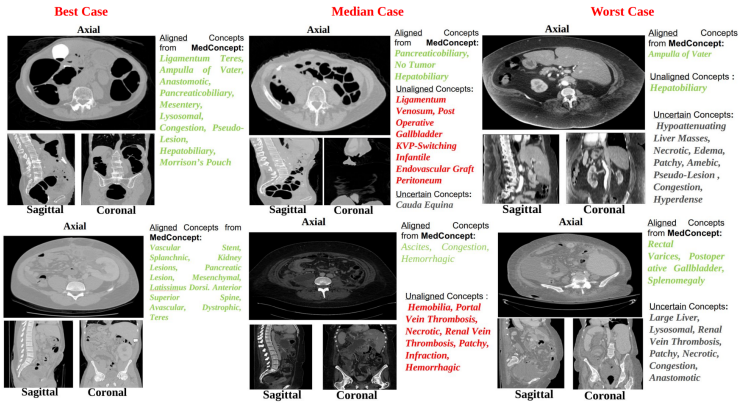

Fig. 2: **Qualitative evaluation of unsupervised 3D concept extraction using** MedConcept. The top ten activated concepts per case are displayed. Concepts are evaluated post hoc for clinical grounding via *MedGemma*, which classifies each concept as *Present* (aligned), *Absent* (unaligned or contradicted), or *Uncertain*. Results indicate coherent sparse decomposition and consistent semantic grounding across datasets

rare findings. Because evaluation is conditioned on report evidence, such labels may reflect model overreach or simply incomplete documentation, rather than definitive clinical incorrectness. In the *lowest-scoring cases*, predictions are predominantly Uncertain rather than explicitly Unaligned, often involving generic imaging descriptors (e.g., hypoattenuating, patchy, necrotic, hyperdense) that are difficult to verify against incomplete reports. This progression from Aligned to increasingly Unaligned and Uncertain predictions mirrors the quantitative trends in verification scores, indicating that performance degradation reflects semantic overreach or report incompleteness rather than anatomically implausible predictions. The progression from predominantly Aligned predictions to regimes characterized by increased Unaligned (contradiction-driven) and subsequently Uncertain (ambiguity-driven) concepts mirrors the quantitative trends in verification scores. Collectively, these examples demonstrate that the framework produces structured and interpretable concepts when supported by sufficient textual evidence, while performance degradation reflects semantic overreach or report incompleteness rather than arbitrary or anatomically implausible predictions.

Quantitative Analysis: As shown in Fig. 3, concept verification scores exhibit consistent and interpretable patterns across datasets and settings. Distributions

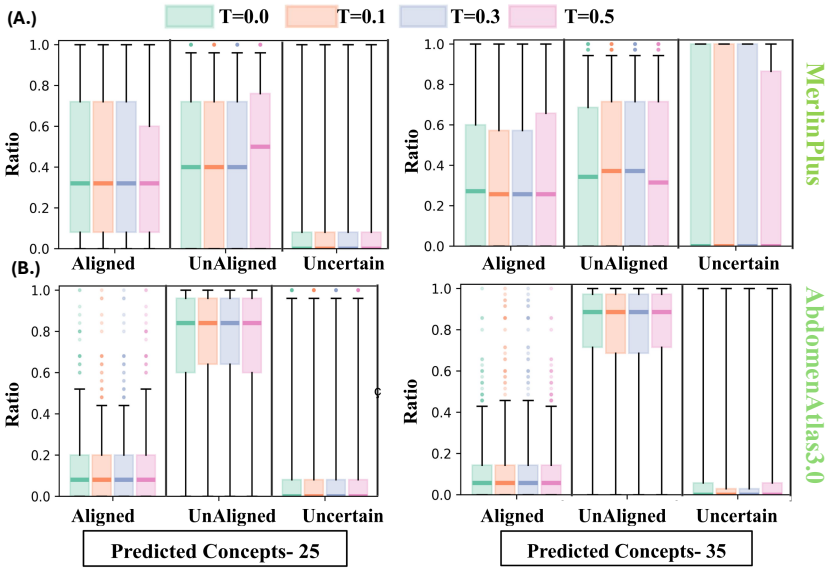

Fig. 3: **Boxplots of concept verification scores for the top-K predicted concepts ($K \in 25, 35$).** Each panel shows the distribution of per-volume **Concept Alignment** (Aligned), **Concept UnAlignment** (UnAligned), and **Concept Uncertain** scores aggregated across cases. Alignment scores are consistently higher on the (A.) MerlinPlus dataset compared to (B.) Abdomen Atlas, while Abdomen Atlas exhibits comparatively lower alignment and higher unalignment proportions across temperature settings.

remain nearly invariant across LLM decoding temperatures from 0.0 to 0.5, indicating that the scoring procedure is stable and not driven by sampling variability under fixed prompts. On MerlinPlus, the top 25 predicted concepts achieve a median Aligned score of approximately 0.30–0.35, while Uncertain scores typically remain below 0.1. Expanding to the top 35 concepts results in a moderate decrease in alignment and a corresponding increase in uncertainty. This rank-dependent shift is consistent with the inclusion of lower-activation units and suggests an implicit ordering within the sparse representation, where higher-activation units tend to correspond to more semantically supported concepts. The coherent behavior across ranks indicates structured semantic organization within the learned volumetric representations and meaningful concept assignment through cosine-based retrieval over the curated medical vocabulary. In contrast, AbdomenAtlas 3.0 exhibits substantially lower alignment, with median Aligned scores below 0.1 and Unaligned scores above 0.8 for the top 25 concepts. The top 35 configuration shows a similar distribution with only a modest increase in uncertainty. These patterns remain stable across temperature settings. The reduced alignment on this dataset is likely influenced by differences in report density and documentation style, which constrain textual evidence available for semantic verification.

Temperature and λ Ablation Results. Furthermore, across temperatures $T \in 0.1, 0.3, 0.5$, $\lambda_1 = 2 \times 10^{-3}$ provided the most favorable selectivity–coverage trade-off, whereas strong sparsity ($\lambda_1 = 0.2$) suppressed concept activation and weak sparsity ($\lambda_1 = 2 \times 10^{-4}$) increased activation coverage but produced noisier and more variable report-grounded concepts.

Limitations. While Figures 2 and 3 illustrate structured concept discovery without manual annotations, quantitative alignment scores are inherently bounded by the characteristics of the paired radiology reports used for evaluation. Because semantic verification is conditioned on report evidence, concepts visually present in the CT volume but omitted from documentation are penalized as Unaligned. This limitation reflects the selective and non-exhaustive nature of clinical reporting rather than definitive semantic error. Evaluation is further constrained by vocabulary coverage. Although curated from UMLS and refined via LLM expansion, the predefined concept set cannot exhaustively represent all abdominal pathologies, anatomical variants, or contextual findings. Beyond textual constraints, the framework has an algorithmic limitation: semantic naming relies on cosine similarity within the joint VLM latent space and therefore inherits representational biases and potential vision–text misalignments present in the underlying pretrained model.

4 Conclusion and Future Work

We introduce MedConcept, a unified framework for unsupervised concept extraction and structured evaluation in medical Vision–Language Models. To the best of our knowledge, this study is among the first to systematically extract concepts from 3D volumetric foundation model representations and to formalize

their quantitative verification. We apply sparse autoencoders to frozen volumetric embeddings to decompose high-dimensional features into sparse neuron-level activations, grounding them in medical terminology through a structured vocabulary within a shared vision–language embedding space. A central contribution is a scalable LLM-based evaluation protocol that defines Aligned, Unaligned, and Uncertain scores to quantify concept–report consistency in a reproducible and temperature-robust manner. This framework enables structured verification across cases and concept ranks, transforming concept interpretability from qualitative inspection into a measurable and comparable evaluation setting. Experiments on AbdomenAtlas 3.0 and MerlinPlus demonstrate stable scoring across decoding temperatures and consistent rank-dependent trends, highlighting robustness to LLM stochasticity.

Future work will focus on expanding and refining the medical vocabulary through expert curation, conducting systematic validation with clinician-annotated studies, extending evaluation to additional datasets and foundation models, and developing methods for improved concept disentanglement and hierarchical organization within volumetric representations.

Acknowledgments. The authors acknowledge the support of the Kahlert School of Computing and the Scientific Computing and Imaging Institute at the University of Utah.

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

References

1. Bassi, P.R., Li, W., Chen, J., Zhu, Z., Lin, T., Decherchi, S., Cavalli, A., Wang, K., Yang, Y., Yuille, A.L., et al.: Learning segmentation from radiology reports. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 305–315. Springer (2025)
2. Bhalla, U., Oesterling, A., Srinivas, S., Calmon, F., Lakkaraju, H.: Interpreting clip with sparse linear concept embeddings (splice). Advances in Neural Information Processing Systems **37**, 84298–84328 (2024)
3. Blankemeier, L., Cohen, J.P., Kumar, A., Van Veen, D., Gardezi, S.J.S., Paschali, M., Chen, Z., Delbrouck, J.B., Reis, E., Truyts, C., et al.: Merlin: A vision language foundation model for 3d computed tomography. Research Square pp. rs–3 (2024)
4. Bodenreider, O.: The unified medical language system (umls): integrating biomedical terminology. Nucleic acids research **32**(suppl_1), D267–D270 (2004)
5. Bricken, T.: Towards monosemanticity: Decomposing language models with dictionary learning. `https://transformer-circuits.pub/2023/monosemantic-features` (2023), transformer Circuits Thread
6. Chung, M., Won, J.B., Kim, G., Kim, Y., Ozbulak, U.: Evaluating visual explanations of attention maps for transformer-based medical imaging. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 110–120. Springer (2024)
7. Dasdelen, M.F., Lim, H., Buck, M., Götze, K.S., Marr, C., Schneider, S.: Cytosae: Interpretable cell embeddings for hematology. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 77–86 (2025)

8. Dhar, T., Dey, N., Borra, S., Sherratt, R.S.: Challenges of deep learning in medical image analysis—improving explainability and trust. IEEE Transactions on Technology and Society **4**(1), 68–75 (2023)
9. D'Antonoli, T.A., Bluethgen, C., Cuocolo, R., Klontzas, M.E., Ponsiglione, A., Kocak, B.: Foundation models for radiology: fundamentals, applications, opportunities, challenges, risks, and prospects. Diagnostic and Interventional Radiology (2025)
10. Ennab, M., Mcheick, H.: Enhancing interpretability and accuracy of ai models in healthcare: a comprehensive review on challenges and future directions. Frontiers in Robotics and AI **11**, 1444763 (2024)
11. Ennab, M., Mcheick, H.: Advancing ai interpretability in medical imaging: a comparative analysis of pixel-level interpretability and grad-cam models. Machine Learning and Knowledge Extraction **7**(1), 12 (2025)
12. Gallifant, J., Chen, S., Sasse, K., Aerts, H., Hartvigsen, T., Bitterman, D.: Sparse autoencoder features for classifications and transferability. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. pp. 29927–29951 (2025)
13. Gong, S., Wang, H., Zhang, X., Dou, Q.: Concepts from neurons: Building interpretable medical image diagnostic models by dissecting opaque neural networks. In: International Conference on Information Processing in Medical Imaging. pp. 3–18. Springer (2025)
14. Komorowski, P., Baniecki, H., Biecek, P.: Towards evaluating explanations of vision transformers for medical imaging. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3726–3732 (2023)
15. Lepcha, D.C., Goyal, B., Dogra, A., Alkhayyat, A., Sahu, P.K., Ali, A., Kukreja, V.: Deep learning in medical image analysis: A comprehensive review of algorithms, trends, applications, and challenges. Computer Modeling in Engineering & Sciences **145**(2), 1487 (2025)
16. Li, W., Qu, C., Chen, X., Bassi, P.R., Shi, Y., Lai, Y., Yu, Q., Xue, H., Chen, Y., Lin, X., et al.: Abdomenatlas: A large-scale, detailed-annotated, & multi-center dataset for efficient transfer learning and open algorithmic benchmarking. Medical Image Analysis **97**, 103285 (2024)
17. Lindberg, D.A., Humphreys, B.L., McCray, A.T.: The unified medical language system. Yearbook of medical informatics **2**(01), 41–51 (1993)
18. OpenAI: Chatgpt (5.2) (2026), `https://chat.openai.com/chat`
19. Rao, S., Mahajan, S., Böhle, M., Schiele, B.: Discover-then-name: Task-agnostic concept bottlenecks via automated concept discovery. In: European Conference on Computer Vision. pp. 444–461. Springer (2024)
20. Salahuddin, Z., Woodruff, H.C., Chatterjee, A., Lambin, P.: Transparency of deep neural networks for medical image analysis: A review of interpretability methods. Computers in biology and medicine **140**, 105111 (2022)
21. Sellergren, A., Kazemzadeh, S., Jaroensri, T., Kiraly, A., Traverse, M., Kohlberger, T., Xu, S., Jamil, F., Hughes, C., Lau, C., et al.: Medgemma technical report. arXiv preprint arXiv:2507.05201 (2025)
22. Shi, W., Li, S., Liang, T., Wan, M., Ma, G., Wang, X., He, X.: Route sparse autoencoder to interpret large language models. In: Christodoulopoulos, C., Chakraborty, T., Rose, C., Peng, V. (eds.) Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. pp. 6801–6815. Association for Computational Linguistics (Nov 2025)

23. Wu, J., Wang, Y., Zhong, Z., Liao, W., Trayanova, N., Jiao, Z., Bai, H.X.: Vision-language foundation model for 3d medical imaging. npj Artificial Intelligence **1**(1), 17 (2025)