Abnormality-Driven Representation Learning for Radiology Imaging

Marta Ligero[1†], Tim Lenz[1†], Georg Wölflein[1,2], Omar S.M. El Nahhas[1,3], Daniel Truhn[4], and Jakob Nikolas Kather[1,3,4,5,6*]

[1] EKFZ for Digital Health TU Dresden, Germany
[2] University of St Andrews, United Kingdom
[3] StratifAI GmbH, Germany
[4] RWTH Aachen University, Germany
[5] Heidelberg University Hospital, Germany
[6] University Hospital Dresden, Germany
jakob_nikolas.kather@tu-dresden.de

Abstract. Radiology deep learning pipelines predominantly employ end-to-end 3D networks based on models pre-trained on other tasks, which are then fine-tuned on the task at hand. In contrast, adjacent medical fields such as pathology, which focus on 2D images, have effectively adopted task-agnostic foundational models based on self-supervised learning (SSL), combined with weakly-supervised deep learning (DL). However, the field of radiology still lacks task-agnostic representation models due to the computational and data demands of 3D imaging and the anatomical complexity inherent to radiology scans. To address this gap, we propose CLEAR, a framework for 3D radiology images that uses extracted embeddings from 2D slices along with attention-based aggregation to efficiently predict clinical endpoints. As part of this framework, we introduce LECL, a novel approach to obtain visual representations driven by abnormalities in 2D axial slices across different locations of the CT scans. Specifically, we trained single-domain contrastive learning approaches using three different architectures: Vision Transformers, Vision State Space Models and Gated Convolutional Neural Networks. We evaluate our approach across three clinical tasks: tumor lesion location, lung disease detection, and patient staging, benchmarking against four state-of-the-art foundation models, including BiomedCLIP. Our findings demonstrate that CLEAR, using representations learned through LECL, outperforms existing foundation models, while being substantially more compute- and data-efficient. The code is available at https://github.com/KatherLab/CLEAR.

1 Introduction

Recent advances in precision oncology have highlighted the need for artificial intelligence (AI) systems capable of analyzing whole-body radiology images to

[†] Equal contribution, * Corresponding author

characterize metastatic cancer patients [24]. The development of spatial biomarkers from radiology predominantly consists of the implementation of either handcrafted features (radiomics), or end-to-end deep learning (DL) pipelines (Fig. 1A) [5]. These approaches, however, both require manual or automated scan selection, followed by identification and manual annotation of lesions [23]. With DL systems demanding extensive annotated datasets for specific tasks, these resource-intensive processing requirements represent a substantial bottleneck in current biomarker development pipelines [5, 23].

In contrast, adjacent fields such as pathology has established effective pipelines using task-agnostic foundation models combined with attention-based aggregation, reducing the need for extensive preprocessing [9]. However, radiology lacks such foundation models that can extract generalizable imaging representations without task-specific fine-tuning of the encoder [22]. Existing models are either too specific to certain regions and applications [21, 6, 10] or overly generalized across modalities [29].

To address these challenges, our contributions are as follows:

- We propose Contrastive Learning-based Embeddings for Attention-based Radiology (**CLEAR**) (see Fig. 1B), a domain-specific framework for radiology that integrates foundation models and attention methods for model development in diverse clinical applications.
- As part of this framework, we introduce Lesion-enhanced Contrastive Learning (**LeCL**) (Fig. 1C), a novel semi-supervised method, and compare it with MoCo-v3 [7] for feature representation of abnormal lesions throughout the whole body.
- We performed a comprehensive analysis of different 2D-based model architectures, including Vision Transformers (ViT), Vision State Space Models (VSSM) and gated Convolutional Neural Networks (CNN), to develop more effective domain-specific foundation models.

2 Related work

Most research in radiology for developing non-invasive biomarkers has relied on radiomics [1]. Although recent deep learning advancements have enabled direct prediction of treatment response without manual feature extraction [16–18], these approaches require fine-tuning. Inspired by the success in the pathology field where learned representations are combined with attention-based methods, we propose a framework that integrates frozen pretrained features with attention-based multiple instance learning (MIL) for radiology.

Progress has been made towards foundation models for imaging biomarkers using self-supervised learning in radiology [13]. Initially targeting radiographs [4], recent work has expanded to 3D modalities such as CT and MRI [26, 10, 6]. However, the largest foundation models remain proprietary [28], while

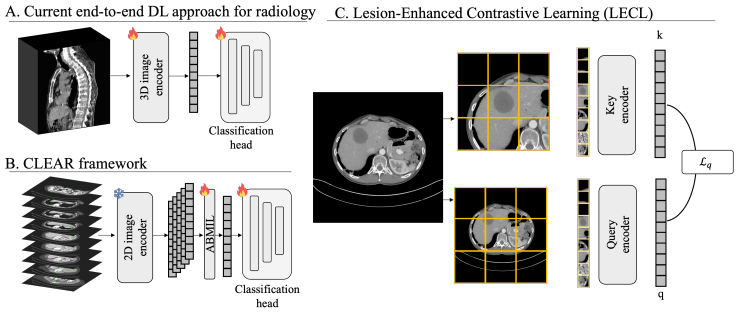

Fig. 1: **Overview of the proposed framework CLEAR**. Currently, end-to-end deep learning approaches in radiology mostly fine-tune the encoder for each specific task separately (A). We developed a weakly supervised pipeline that deploys a pretrained encoder to extract frozen embeddings, which are used for supervised training of an attention-based pooling model (B). For pretraining the feature extractor, we introduce LECL, a semi-supervised algorithm that guarantees that the abnormalities are within the crops of the images (C).

other approaches require lesion annotations during inference [21]. We proposed a novel, open-source SSL foundation model for CT scans with a lesion-aware semi-supervised framework utilizing specific cropping strategies during training, similar to successful approaches in other medical imaging fields [14].

While existing methods often rely on computationally intensive 3D architectures requiring encoder fine-tuning [10, 6], attention-based approaches have shown promise in capturing complex imaging patterns [3, 19, 12]. Our study introduces a computationally efficient approach that combines 2D slice-based representation with attention mechanisms, outperforming other 3D SSL models [10, 6] and achieving comparable performance to general-purpose models like BiomedCLIP [29] despite training on a smaller dataset, while eliminating the need for lesion-level annotations at inference time.

3 Methods

As part of the CLEAR framework, we introduce LECL, a semi-supervised DL framework for representation learning from radiology images, building upon contrastive SSL methods. Our method uses 2D axial CT scan slices.

3.1 Lesion enhanced contrastive learning

To increase the focus on lesions in CT scans, we propose LECL. This method ensures that the momentum encoder receives lesion-centered image crops for annotated slices, while the key encoder embeds the full image of the same slice. Let $\boldsymbol{q}$ and $\boldsymbol{k}$ denote the query and key feature vectors, respectively, derived from

different random augmentations of the same input image. The InfoNCE loss function [20] is given by:

$$\mathcal{L}_{\mathbf{q}} = -\log \frac{\psi(\mathbf{q}, \mathbf{k}^+)}{\sum_{i=1}^{N} \psi(\mathbf{q}, \mathbf{k}_i)}, \tag{1}$$

where $\boldsymbol{k}^+$ represents the positive key (i.e., the key corresponding to the same image as the query $\boldsymbol{q}$), $\boldsymbol{k}_i$ represents the i-th negative key, and $\psi(\mathbf{x}_1, \mathbf{x}_2) = \exp(\text{sim}(\mathbf{x}_1, \mathbf{x}_2)/\tau)$, with τ representing the temperature parameter and $\text{sim}(\cdot)$ denoting cosine similarity. We introduce an additional term $\xi = \sum_{j=1}^{L} \psi(\mathbf{q}, \mathbf{l_j})$, with $\{l_j\}_{j\in\{1,\ldots,L\}}$ denoting the set of key slices of the training set, to increase the weight of key slice encodings:

$$\mathcal{L}_{\mathbf{q}}^{\text{LECL}} = -\log \frac{\psi(\mathbf{q}, \mathbf{k}^+)}{\sum_{i=1}^{N} \psi(\mathbf{q}, \mathbf{k}_i) + \lambda \cdot \xi}, \tag{2}$$

where λ denotes a weighting factor for the introduced term. Our implementation is based on the MoCo-v3 repository [7], following the principles of contrastive learning outlined by He et al. [11].

3.2 Weakly supervised learning on frozen features

The embeddings from all axial slices of the CT scan are extracted using contrastive learning methods and serve as input for the subsequent classification module. The aggregation of slice embeddings H is defined by the MIL pooling function $f : \mathbb{R}^{K\times d} \to \mathbb{R}^d$ [15]:

$$\boldsymbol{z} = f(H) = \sum_{k=1}^{K} a_k(\boldsymbol{h}_k) \cdot \boldsymbol{h}_k, \tag{3}$$

where $a_k(\boldsymbol{h}_k) = \frac{\exp\left(\boldsymbol{w}^\top \tanh\left(\boldsymbol{V}\boldsymbol{h}_k^\top\right)\right)}{\sum_i^K \exp\left(\boldsymbol{w}^\top \tanh\left(\boldsymbol{V}\boldsymbol{h}_i^\top\right)\right)}$, with learnable parameters $\boldsymbol{w} \in \mathbb{R}^{p\times 1}$ and $\boldsymbol{V} \in \mathbb{R}^{p\times d}$.

3.3 Cohort description

DeepLesion. We included 14,601 contrast enhanced CT scans from 4,427 patients with solid tumors in different regions including bone, lung, mediastinum, liver, kidney, abdomen, soft tissue and pelvis from the DeepLesion dataset [27]. Lesion bounding boxes were available for all CT scans whereas lesion labels were available for 4,177 CT scans from 1,368 patients. This dataset was used to train representation learning methods as well as for downstream task evaluation. Labeled CT scans were used for predicting lesion location as internal downstream task evaluation (Task 1). We separated a subset of 839 patients (20%) as a test set for evaluation. All 10,224 unlabeled CT scans and 3,538 labeled CT scans (80%) were included for model pre-training. The downstream task was trained on the overlapping labeled CT scans and evaluated on the held out test set.

RAD-ChestCT. We included 3,630 non-contrast enhanced chest CT scans from patients with several abnormalities in the lung including emphysema, bronchiectasis, pleural effusion, consolidation, calcification, bronchial wall thickening, atelectasis, fibrosis, opacity and lung nodules from the RAD-ChestCT dataset [8]. This dataset was used to evaluate the performance of pre-trained models for predicting lung abnormalities as a downstream task (Task 2).

NSCLC-Radiomics. We included 447 non-contrast enhanced CT scans from 422 patients with Non-Small Cell Lung Cancer (NSCLC) tumors from the NSCLC-Radiomics dataset [2] to evaluate the performance of pre-trained models for predicting patient stage defined as low (Stage I and II) and high (Stage III and IV) as a downstream task (Task 3).

3.4 Image preprocessing

We consider each CT scan as a set of axial slices 700 axial slices (512×512 px) with Hounsfield units between -1024 and 1024. To enhance lesion visibility during pre-training, we clip images following DeepLesion guidelines, using abdominal windows (-175 to 275 HU) for general tissue and lung windows (-1500 to 500 HU) for pulmonary structures. Task 1 employs both windows to capture diverse lesions, while Tasks 2 and 3 use only lung windows for their pulmonary focus.

3.5 Experimental setup

Pre-training details We evaluate MoCo and LeCL pretraining across three architectures (ViT, VMamba, and MambaOut), using 873,849 axial CT slices from DeepLesion (see Table 1 for architecture details). Pretraining took less than 2 days on 4 A100 GPUs (ViT-B/VMamba: 33h, MambaOut: 25h) for 100 epochs with batch sizes of 1024-2048 and learning rate 1e$-$4 with 10 warm-up epochs. We tested LeCL with $\lambda \in 0, 1, 3, 5$ (see Eq. (2)) All other parameters followed the official MoCo-v3 repository configuration [7].

Downstream evaluation. We adopted the conventional linear protocol to evaluate features of pretrained models. This approach involves freezing the backbone network weights while training only the subsequent adapter. We evaluated the performance of the frozen embeddings in three different downstream tasks and compared them against four different pre-training models as feature extractors, including both 2D approaches (BiomedCLIP [29] and SAM2 [25]) and 3D approaches (CT-CLIP [10] and Merlin [6]) (see Table 1 for model details). For Tasks 1-2, we applied multi-class multi-label classification with binary cross-entropy loss, while Task 3 used standard classification with cross-entropy loss. All models used learning rate 1e-4, batch size 128, and 32 epochs with early stopping. We employed 5-fold cross-validation with separate test sets for Tasks 1-2 and nested 5-fold cross-validation in Task 3. We evaluated performance using Area Under the Receiver Operating Characteristic (AUC).

Table 1: Overview of the number of parameters per model.

Model	# Parameters [M]
BiomedCLIP [29]	86
Merlin [6]	122
SAM2 [25]	213
CT-CLIP [10]	1110
Ours-VMamba	36
Ours-MambaOut	22
Ours-ViT-B	82

Table 2: **Comparison of different foundation models.** AUC performance of downstream tasks. Internal validation in DeepLesion (Task 1). The mean over five folds is reported alongside the standard deviation as subscript.

Model	Abdomen	Mediastinum	Pelvis	Bone	Soft Tissue	Kidney	Liver	Lung	Average
Merlin[6]	$57.6_{2.4}$	$54.4_{0.2}$	$50.8_{0.6}$	$50.0_{0.0}$	$50.0_{0.0}$	$50.0_{0.0}$	$50.3_{0.1}$	$59.5_{5.1}$	$52.8_{2.0}$
CT-CLIP [10]	$69.1_{0.8}$	$60.5_{2.4}$	$57.2_{1.7}$	$50.0_{0.0}$	$50.7_{0.7}$	$49.9_{0.2}$	$56.7_{1.0}$	$75.7_{0.5}$	$58.7_{1.2}$
SAM2 [25]	$\underline{84.4}_{2.4}$	$89.0_{2.7}$	$86.3_{6.8}$	$54.6_{2.8}$	$64.2_{1.4}$	$53.8_{3.6}$	$77.3_{4.0}$	$88.3_{0.5}$	$74.7_{3.5}$
BiomedCLIP [29]	$\mathbf{85.4}_{1.0}$	$89.3_{1.9}$	$91.8_{2.1}$	$53.3_{4.0}$	$66.0_{1.7}$	$\mathbf{65.9}_{1.7}$	$78.6_{2.4}$	$\underline{90.9}_{0.9}$	$77.6_{2.2}$
MambaOut-MoCo	$81.3_{1.4}$	$88.6_{1.2}$	$\mathbf{94.8}_{0.9}$	$53.3_{2.8}$	$69.5_{0.9}$	$63.9_{2.0}$	$79.9_{2.7}$	$90.1_{0.4}$	$77.7_{1.7}$
MambaOut-LeCL-1	$81.7_{1.6}$	$88.7_{2.2}$	$94.2_{0.9}$	$54.6_{3.8}$	$68.8_{1.5}$	$63.1_{2.0}$	$\underline{81.5}_{1.7}$	$90.3_{0.8}$	$77.9_{2.0}$
MambaOut-LeCL-0	$82.9_{1.3}$	$\underline{89.4}_{1.5}$	$\underline{94.7}_{0.5}$	$\mathbf{\underline{55.4}}_{4.4}$	$69.6_{2.6}$	$\underline{64.0}_{0.5}$	$80.1_{2.2}$	$90.8_{0.5}$	$\mathbf{78.4}_{2.1}$
VMamba-MoCo	$82.4_{2.0}$	$\mathbf{90.5}_{0.3}$	$92.8_{1.4}$	$\mathbf{\underline{55.4}}_{4.5}$	$67.3_{1.3}$	$63.6_{3.4}$	$79.5_{2.8}$	$89.9_{0.4}$	$77.7_{2.4}$
VMamba-LeCL-1	$80.6_{1.2}$	$87.5_{1.8}$	$94.6_{0.6}$	$53.2_{2.7}$	$\underline{71.0}_{2.2}$	$62.3_{2.7}$	$81.3_{2.2}$	$\underline{90.9}_{0.5}$	$77.7_{1.9}$
VMamba-LeCL-0	$82.1_{1.0}$	$88.1_{1.8}$	$94.5_{1.6}$	$54.6_{3.8}$	$\mathbf{71.9}_{2.3}$	$63.9_{2.5}$	$78.9_{2.2}$	$\mathbf{91.2}_{0.5}$	$\underline{78.2}_{2.2}$
ViT-LeCL-0	$81.1_{1.2}$	$88.5_{2.3}$	$91.4_{1.1}$	$50.4_{0.8}$	$67.4_{2.3}$	$61.0_{3.1}$	$79.3_{1.8}$	$89.2_{0.6}$	$76.0_{1.8}$
ViT-LeCL-1	$80.0_{1.6}$	$85.9_{1.5}$	$92.8_{1.0}$	$50.8_{1.0}$	$68.4_{1.6}$	$61.5_{2.2}$	$80.1_{1.2}$	$90.1_{0.2}$	$76.2_{1.4}$
ViT-ConvB	$81.0_{1.9}$	$86.4_{1.1}$	$93.5_{0.5}$	$51.6_{2.0}$	$67.2_{2.9}$	$61.7_{1.2}$	$\mathbf{81.6}_{1.5}$	$89.7_{0.3}$	$76.6_{1.6}$

4 Results

Our findings demonstrate that vision-only representation learning on smaller datasets performs comparably to larger multimodal architectures like BiomedCLIP. Our proposed LeCL approach with MambaOut achieved superior results: +0.8% AUC over BiomedCLIP in lesion classification (Task 1), +1.90% AUC in chest abnormality detection (Task 2), and +3.10% AUC in patient staging (Task 3), with greater gains in specific conditions like soft tissue (+5.90% AUC) and emphysema (+5.8% AUC). More detailed results for each task can be found in Tables 2 to 4. Additionally, our ablation study for different lesion weighting parameters for contrastive learning (Figure 2) shows that LeCL improves performance compared to MoCo in all tasks for MambaOut (+0.5%, +0.35% and +0.35% AUC in Task 1 to 3) and for VMamba (+0.5% AUC in Task 1,2). However, increasing the value of the parameter for weighting lesion representation showed a drop in performance for $\lambda \in \{1, 3, 5\}$. With a maximum drop for $\lambda = 5$ in both MambaOut (up to -1.19% and -4% AUC for Task 1,2) and VMamba (-0.79% and -0.9% AUC).

Table 3: **Comparison of different foundation models.** AUC performance of downstream tasks. External validation on RAD-ChestCT (Task 2).

Model	Emphysema	Bronchiectasis	Pleural Effusion	Atelectasis	Fibrosis	Opacity	Calcification	Lung Nodule	Average
CT-CLIP [10]	$50.6_{0.5}$	$50.0_{0.0}$	$53.2_{1.2}$	$50.8_{0.4}$	$50.0_{0.0}$	$51.6_{1.0}$	$50.8_{0.4}$	$51.6_{0.5}$	$51.1_{0.6}$
SAM2 [25]	$57.6_{1.9}$	$50.0_{0.0}$	$64.6_{3.7}$	$54.6_{2.4}$	$50.4_{0.8}$	$58.4_{1.9}$	$57.8_{2.1}$	$58.0_{2.3}$	$56.4_{2.2}$
Merlin [6]	$64.0_{1.4}$	$52.6_{0.5}$	$67.6_{1.9}$	$57.0_{1.4}$	$56.2_{1.2}$	$59.8_{0.4}$	$\mathbf{65.8}_{1.6}$	$55.6_{1.0}$	$59.8_{1.3}$
BiomedCLIP [29]	$69.8_{1.6}$	$\underline{66.6}_{1.2}$	$75.0_{4.8}$	$60.8_{1.2}$	$66.8_{1.2}$	$60.4_{1.9}$	$62.2_{1.7}$	$58.8_{2.2}$	$65.0_{2.3}$
MambaOut-LeCL-1	$74.0_{1.4}$	$64.6_{1.2}$	$80.2_{1.0}$	$62.2_{1.2}$	$67.6_{1.2}$	$\mathbf{61.2}_{2.3}$	$61.8_{2.7}$	$58.2_{2.8}$	$66.2_{1.9}$
MambaOut-MoCo	$\underline{75.0}_{1.7}$	$66.0_{0.6}$	$\mathbf{81.4}_{1.4}$	$61.4_{1.4}$	$\underline{67.8}_{1.2}$	$59.2_{1.2}$	$62.4_{1.2}$	$59.8_{1.5}$	$\underline{66.6}_{1.3}$
MambaOut-LeCL-0	$\mathbf{75.6}_{1.0}$	$\mathbf{67.6}_{2.7}$	$79.6_{1.0}$	$\underline{63.2}_{2.3}$	$66.8_{1.6}$	$\underline{60.8}_{1.2}$	$61.8_{1.3}$	$59.8_{1.2}$	$\mathbf{66.9}_{1.6}$
VMamba-MoCo	$72.6_{0.5}$	$63.2_{1.3}$	$\underline{80.6}_{1.0}$	$62.2_{0.8}$	$64.0_{2.4}$	$57.2_{2.0}$	$61.8_{1.3}$	$\mathbf{61.8}_{1.9}$	$65.4_{1.5}$
VMamba-LeCL-1	$73.6_{2.0}$	$63.4_{1.7}$	$78.6_{0.5}$	$62.6_{0.5}$	$67.6_{4.6}$	$58.2_{1.2}$	$62.4_{1.4}$	$59.4_{1.6}$	$65.7_{2.1}$
VMamba-LeCL-0	$73.8_{0.8}$	$65.0_{1.1}$	$77.0_{0.6}$	$60.6_{1.2}$	$\mathbf{68.2}_{1.2}$	$59.2_{0.8}$	$\underline{63.2}_{1.2}$	$\underline{60.0}_{1.4}$	$65.9_{1.1}$
ViT-ConvB	$65.4_{3.6}$	$50.4_{0.5}$	$77.2_{1.5}$	$60.2_{0.8}$	$50.6_{0.5}$	$56.6_{2.1}$	$62.2_{1.0}$	$57.6_{1.0}$	$60.0_{1.7}$
ViT-LeCL-0	$73.0_{0.9}$	$59.4_{3.5}$	$76.2_{0.8}$	$62.0_{1.8}$	$66.2_{3.7}$	$59.8_{1.5}$	$62.4_{0.8}$	$59.6_{0.8}$	$64.8_{2.1}$
ViT-LeCL-1	$72.8_{1.0}$	$63.6_{2.8}$	$78.0_{1.8}$	$\mathbf{64.0}_{1.4}$	$64.8_{2.5}$	$56.4_{2.1}$	$62.6_{1.2}$	$58.8_{1.3}$	$65.1_{1.9}$

Table 4: **Comparison of different foundation models.** Performance of downstream tasks. External validation on NSCLC-Radiomics (Task 3).

model	AUC	AUPRC	F1
CT-CLIP [10]	$51.6_{2.8}$	$69.3_{0.8}$	$63.8_{22.3}$
Merlin [6]	$60.9_{5.7}$	$74.1_{3.4}$	$67.2_{5.3}$
SAM2 [25]	$61.4_{5.1}$	$74.3_{2.8}$	$69.8_{5.5}$
BiomedCLIP [29]	$65.2_{4.2}$	$76.5_{2.6}$	$67.2_{5.4}$
MambaOut-LeCL-1	$64.9_{5.3}$	$76.3_{2.7}$	$70.7_{4.8}$
MambaOut-MoCo	$66.7_{4.1}$	$77.3_{2.3}$	$70.8_{3.7}$
MambaOut-LeCL-0	$\mathbf{68.3}_{5.0}$	$\mathbf{78.3}_{2.7}$	$\mathbf{72.3}_{5.3}$
VMamba-LeCL-1	$64.4_{5.7}$	$76.1_{3.4}$	$69.2_{4.1}$
VMamba-LeCL-0	$65.3_{6.3}$	$76.7_{3.5}$	$69.9_{3.7}$
VMamba-MoCo	$65.5_{9.0}$	$77.0_{5.1}$	$70.4_{5.9}$
ViT-LeCL-0	$61.5_{7.6}$	$74.7_{4.0}$	$66.6_{5.0}$
ViT-LeCL-1	$64.8_{4.9}$	$76.2_{2.7}$	$69.8_{4.2}$
ViT-ConvB	$\underline{67.5}_{3.7}$	$\underline{77.8}_{2.0}$	$\underline{71.5}_{5.9}$

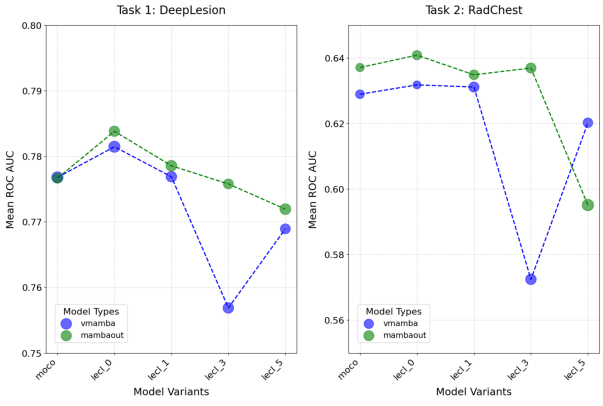

Fig. 2: **Contrastive lesion weight ablation**. AUC comparison across tasks for hyperparameter λ (see Eq. (2)).

Finally, Fig. 3 shows the model highlighting the most informative slices from the CT scans with higher attention while assigning less attention to images with healthy tissues or uninformative CT scan acquisitions. These findings indicate that our framework streamlines image processing by eliminating manual preprocessing selection and tumor annotation steps by using attention mechanisms.

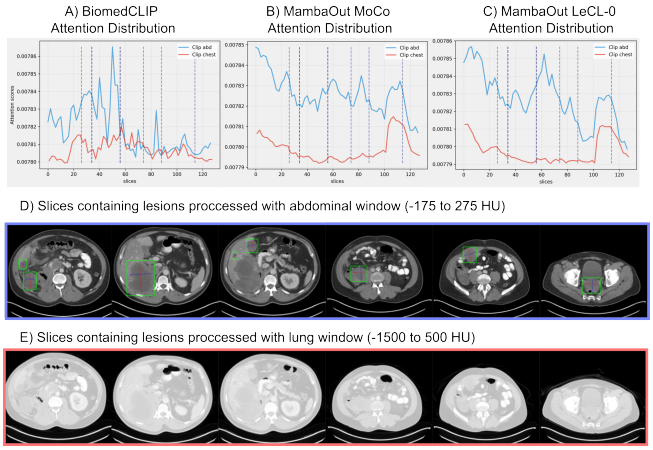

Fig. 3: **Attention distribution across different slices**: We evaluated the attention distribution across slices for a patient with liver and soft tissue lesions for BiomedCLIP (A), MambaOut architecture trained using MoCo (B) and MambaOut using LECL approach for $\lambda = 0$ (C). Blue represents attention for slices processed in abdominal window images (D) and red represents slices processed in lung window (E). All models show higher attention to the abdominal window where the lesion is better depicted.

5 Conclusion

We introduced CLEAR, a novel framework for radiology image classification based on representation learning. Inspired by the success of attention-based methods in pathology, our framework combines frozen embeddings with weakly supervised deep learning, showing improved performance while reducing the need for manual annotations. Within this framework, we proposed LECL as a method to learn lesion-aware representations. Our analysis revealed substantial limitations in current models for image representation, highlighting the need for more domain-specific models using representation learning approaches. By demonstrating the effectiveness of using frozen embeddings from foundation models, we provide a practical and efficient solution that enables faster development of accurate and reliable radiology image analysis tools.

Acknowledgments. The authors gratefully acknowledge the GWK's support for funding this project by providing computing time through the ZIH at TU Dresden. We also gratefully acknowledge the Gauss Centre for Supercomputing e.V. for funding this project by providing computing time through the NIC on the GCS Supercomputer JUWELS at Jülich Supercomputing Centre. This research was supported by publicly available datasets. We thank the NIH Clinical Center for providing the DeepLesion dataset [27]. We acknowledge the use of RAD-ChestCT dataset [8] under the CC BY-NC-ND 4.0 License, and the NSCLC-Radiomics dataset [2] under the CC BY-NC 3.0 License. JNK is supported by the German Cancer Aid, the German Federal Ministry of Education and Research, the German Academic Exchange Service, the German Federal Joint Committee, the European Union's Horizon Europe research and innovation program, the European Research Council, the National Institutes of Health and the NIHR Leeds Biomedical Research Centre.

Disclosure of Interests. TL declares consulting services for StratifAI. GW declares consulting services for Synagen. OSMEN holds shares in StratifAI. DT holds shares in StratifAI. JNK declares consulting services for Bioptimus; Panakeia; AstraZeneca; and MultiplexDx. Furthermore, he holds shares in StratifAI, Synagen, Tremont AI and Ignition Labs; received an institutional research grant by GSK; and received honoraria by AstraZeneca, Bayer, Daiichi Sankyo, Eisai, Janssen, Merck, MSD, BMS, Roche, Pfizer, and Fresenius.

References

1. Aerts, H., Velazquez, E., Leijenaar, R., et al.: Decoding tumour phenotype by non-invasive imaging using a quantitative radiomics approach. Nature Communications **5**, 4006 (2014). https://doi.org/10.1038/ncomms5006
2. Aerts, H.J.W.L., Wee, L., Velazquez, R., et al.: Data from nsclc-radiomics (2014). https://doi.org/10.7937/K9/TCIA.2015.PF0M9REI
3. Alp, S., Akan, T., Bhuiyan, M., et al.: Joint transformer architecture in brain 3d mri classification: its application in alzheimer's disease classification. Scientific Reports **14**, 8996 (2024). https://doi.org/10.1038/s41598-024-59578-3
4. Bannur, S., Bouzid, K., Castro, D.C., Schwaighofer, A., Thieme, A., Bond-Taylor, S., Ilse, M., Pérez-García, F., Salvatelli, V., Sharma, H., Meissen, F., Ranjit, M., Srivastav, S., Gong, J., Codella, N.C.F., Falck, F., Oktay, O., Lungren, M.P., Wetscherek, M.T., Alvarez-Valle, J., Hyland, S.L.: Maira-2: Grounded radiology report generation (2024), https://arxiv.org/abs/2406.04449
5. Bera, K., Braman, N., Gupta, A., et al.: Predicting cancer outcomes with radiomics and artificial intelligence in radiology. Nature Reviews Clinical Oncology **19**, 132–146 (2022). https://doi.org/10.1038/s41571-021-00560-7
6. Blankemeier, L., Cohen, J.P., Kumar, A., et al.: Merlin: A vision language foundation model for 3d computed tomography (2024), https://arxiv.org/abs/2406.06512
7. Chen*, X., Xie*, S., He, K.: An empirical study of training self-supervised vision transformers. arXiv preprint arXiv:2104.02057 (2021)
8. Draelos, R.L., Dov, D., Mazurowski, M.A., et al.: Machine-learning-based multiple abnormality prediction with large-scale chest computed tomography volumes. Medical Image Analysis **67**, 101857 (2021). https://doi.org/10.1016/j.media.2020.101857

9. El Nahhas, O.S.M., van Treeck, M., Wölflein, G., et al.: From whole-slide image to biomarker prediction: end-to-end weakly supervised deep learning in computational pathology. Nature Protocols (September 2024). https://doi.org/10.1038/s41596-024-01047-2
10. Hamamci, I.E., Er, S., Almas, F., et al.: Developing generalist foundation models from a multimodal dataset for 3d computed tomography (2024), https://arxiv.org/abs/2403.17834
11. He, K., Fan, H., Wu, Y., et al.: Momentum contrast for unsupervised visual representation learning. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9726–9735 (2020). https://doi.org/10.1109/CVPR42600.2020.00975
12. Hu, P., Liu, Y., Li, Y., Zhang, F., Wu, J., Geng, L., Xiao, Z.: Inter-slice attention transformer for predicting risk level of gastrointestinal stromal tumors. In: Proceedings of the 2024 10th International Conference on Computing and Artificial Intelligence. p. 250–259. ICCAI '24, Association for Computing Machinery, New York, NY, USA (2024). https://doi.org/10.1145/3669754.3669791, https://doi.org/10.1145/3669754.3669791
13. Huang, S.C., Pareek, A., Jensen, M., et al.: Self-supervised learning for medical image classification: a systematic review and implementation guidelines. npj Digital Medicine **6**(1), 74 (2023). https://doi.org/10.1038/s41746-023-00811-0
14. Huang, Y., Lin, L., Cheng, P., Lyu, J., Tang, X.: Lesion-based contrastive learning for diabetic retinopathy grading from fundus images (2021), https://arxiv.org/abs/2107.08274
15. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: Dy, J., Krause, A. (eds.) Proceedings of the 35th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 80, pp. 2127–2136. PMLR (10–15 Jul 2018), https://proceedings.mlr.press/v80/ilse18a.html
16. Jiang, X., Zhao, H., Saldanha, O.L., et al.: An mri deep learning model predicts outcome in rectal cancer. Radiology **307**(5), e222223 (2023). https://doi.org/10.1148/radiol.222223, pMID: 37278629
17. Jin, C., Yu, H., Ke, J., et al.: Predicting treatment response from longitudinal images using multi-task deep learning. Nature Communications **12**(1), 1851 (2021). https://doi.org/10.1038/s41467-021-22188-y
18. Lao, J., Chen, Y., Li, Z.C., et al.: A deep learning-based radiomics model for prediction of survival in glioblastoma multiforme. Scientific Reports **7**, 10353 (2017). https://doi.org/10.1038/s41598-017-10649-8
19. Müller-Franzes, G., Khader, F., Siepmann, R., Han, T., Kather, J.N., Nebelung, S., Truhn, D.: Medical slice transformer: Improved diagnosis and explainability on 3d medical images with dinov2 (2024), https://arxiv.org/abs/2411.15802
20. van den Oord, A., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding (2019)
21. Pai, S., Bontempi, D., Hadzic, I., et al.: Foundation model for cancer imaging biomarkers. Nature Machine Intelligence **6**, 354–367 (2024). https://doi.org/10.1038/s42256-024-00807-9
22. Paschali, M., Chen, Z., Blankemeier, L., Varma, M., Youssef, A., Bluethgen, C., Langlotz, C., Gatidis, S., Chaudhari, A.: Foundation models in radiology: What, how, why, and why not. Radiology **314**(2), e240597 (2025). https://doi.org/10.1148/radiol.240597, https://doi.org/10.1148/radiol.240597, pMID: 39903075

23. Perez-Lopez, R., Ghaffari Laleh, N., Mahmood, F., Kather, J.N.: A guide to artificial intelligence for cancer researchers. Nature Reviews Cancer **24**, 427–441 (2024). https://doi.org/10.1038/s41568-024-00694-7
24. Prelaj, A., et al.: Artificial intelligence for predictive biomarker discovery in immuno-oncology: a systematic review. Annals of Oncology **35**(1), 29–65 (2024)
25. Ravi, N., Gabeur, V., Hu, Y.T., et al.: Sam 2: Segment anything in images and videos (2024), https://arxiv.org/abs/2408.00714
26. Wolf, D., Payer, T., Lisson, C.S., Lisson, C.G., Beer, M., Götz, M., Ropinski, T.: Less is more: Selective reduction of ct data for self-supervised pre-training of deep learning models with contrastive learning improves downstream classification performance. Computers in Biology and Medicine **183**, 109242 (Dec 2024). https://doi.org/10.1016/j.compbiomed.2024.109242
27. Yan, K., Wang, X., Lu, L., et al.: Deeplesion: Automated mining of large-scale lesion annotations and universal lesion detection with deep learning. Journal of Medical Imaging **5**(3), 036501 (2018). https://doi.org/10.1117/1.JMI.5.3.036501
28. Yang, L., Xu, S., Sellergren, A., et al.: Advancing multimodal medical capabilities of gemini (2024), https://arxiv.org/abs/2405.03162
29. Zhang, S., Xu, Y., Usuyama, N., et al.: Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs (2024), https://arxiv.org/abs/2303.00915