Computers in Biology and Medicine 195 (2025) 110583

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Radio DINO: A foundation model for advanced radiomics and AI-driven medical imaging analysis

Luca Zedda *, Andrea Loddo, Cecilia Di Ruberto

Department of Mathematics and Computer Science, University of Cagliari, Via Ospedale 72, 09124, Cagliari, Italy

ARTICLE INFO

Keywords:
Radiomics
Self-supervised learning
Deep learning
DINO
DINOV2
Medical imaging
Feature extraction
Generalizability

ABSTRACT

Radiomics is transforming medical imaging by extracting complex features that enhance disease diagnosis, prognosis, and treatment evaluation. However, traditional approaches face significant challenges, such as the need for manual feature engineering, high dimensionality, and limited sample sizes. This paper presents Radio DINO, a novel family of deep learning foundation models that leverage self-supervised learning (SSL) techniques from DINO and DINOV2, pretrained on the RadImageNet dataset. The novelty of our approach lies in (1) developing Radio DINO to capture rich semantic embeddings, enabling robust feature extraction without manual intervention, (2) demonstrating superior performance across various clinical tasks on the MedMNISTv2 dataset, surpassing existing models, and (3) enhancing the interpretability of the model by providing visualizations that highlight its focus on clinically relevant image regions. Our results show that Radio DINO has the potential to democratize advanced radiomics tools, making them accessible to healthcare institutions with limited resources and ultimately improving diagnostic and prognostic outcomes in radiology.

1. Introduction

Radiomics is an emerging field focusing on the comprehensive quantification of medical images, extracting large amounts of advanced imaging features that can provide insights into the underlying tumor biology and characteristics [1]. The promise of radiomics lies in its ability to uncover imaging biomarkers that can complement or even outperform traditional clinical and pathological assessments for various clinical applications, such as disease diagnosis, prognosis prediction, and treatment response evaluation [2–4].

Despite its potential, radiomics faces several critical challenges. Existing radiomic feature extraction approaches often depend on labor-intensive, hand-crafted feature engineering, which may fail to fully capture the complexity and rich information embedded in medical images [5]. Additionally, the high dimensionality and complexity of radiomic data, combined with typically small sample sizes in medical imaging studies, make it difficult to develop robust, generalizable models [6].

To address these limitations, we introduce Radio DINO, a novel family of foundation models designed specifically for radiomics. These models leverage self-supervised learning techniques based on DINO [7] and DINOV2 [8,9] and are pretrained on the diverse RadImageNet dataset. This pre-training enables Radio DINO to capture rich, high-level semantic representations across a wide variety of anatomical regions, organs, and imaging modalities, providing a robust foundation for radiomic feature extraction without the need for manual feature engineering.

Our contributions are threefold. First, we propose Radio DINO, a novel and customized family of deep learning foundation models designed to address traditional radiomics' limitations through self-supervised learning and extensive pre-training on a large radiomics-specific dataset. Second, we demonstrate that Radio DINO achieves superior performance across various clinical tasks, based on heterogeneous data sources like X-ray, ultrasound, and computed tomography images, on the MedMNISTv2 dataset, surpassing state-of-the-art models in several reference performance metrics and generalizability. Third, we show through visualization techniques that Radio DINO is able to focus on clinically relevant image regions, enhancing its interpretability and making it a reliable tool for clinical decision-making.

The implications of our work extend beyond the technical advancements in model training. By providing a powerful, effective foundation model that sets new benchmarks for performance, we aim to make advanced radiomics tools more accessible to a broader range of healthcare institutions, particularly those with limited computational resources. The deployment of Radio DINO could democratize access to sophisticated radiomic analysis, ultimately contributing to improved diagnostic and prognostic outcomes in radiology.

Our proposed models address a significant issue in the radiomics field: the lack of generalizable, radiomics-specific models. In radiomics,

* Corresponding author.
E-mail address: luca.zedda@unica.it (L. Zedda).

https://doi.org/10.1016/j.compbiomed.2025.110583
Received 25 October 2024; Received in revised form 24 May 2025; Accepted 13 June 2025
Available online 28 June 2025

deep learning approaches typically rely on pretrained architectures originally designed for natural images, using supervised or self-supervised methods. Although these methods often yield high performance, they can be overly variable when applied directly to downstream tasks. In contrast, our models consistently deliver high performance, both as standalone feature extractors and as backbones for fine-tuning.

The paper is organized as follows. In Section 2, we review the current state-of-the-art in self-supervised learning and its relevance to medical imaging. Section 3 provides a detailed overview of the datasets and methodologies employed in our experiments. The complete experimental setup is described in Section 4. We present the results of our model evaluations in Section 5, highlighting Radio DINO's performance across various clinical tasks. In Section 6, we explore explainability techniques and feature visualizations to ensure fair and interpretable model predictions. Next, in Section 7, we introduce our proposed evaluation methodologies designed to assess the capabilities of foundation models. Sections 8 and 9 discuss the limitations of our study and provide an in-depth analysis of our findings. To illustrate practical applications, Section 10 outlines how our proposed foundation models can be integrated into real-world clinical workflows. Finally, Section 11 summarizes our key findings and their implications for future research and clinical practice.

2. State of the art

The rapid growth in data acquisition across various fields has resulted in an exponential increase in the size of datasets, driving the need for novel techniques to extract meaningful insights. A key challenge is that much of this data lacks direct annotations, prompting the development of Self-Supervised Learning (SSL) [7,10–13]. SSL has enabled the creation of foundation models capable of learning from unlabeled data, which is particularly useful in fields like medical imaging, where annotated data is scarce.

2.1. Specialized architectures for radiology

Radiology is a high-throughput field that requires domain experts to analyze large volumes of imaging-based reports. To address this challenge, numerous deep learning architectures have been developed, each tailored to specific application needs [14]. Notably, lightweight models with fewer parameters are critical for deployment in resource-constrained environments. This requirement was especially evident during the COVID-19 pandemic, where daily imaging volumes surged and automated feedback became essential [15,16]. Recently, transformer-based architectures have demonstrated remarkable effectiveness across various radiological tasks [17–20]. However, pure transformer models typically suffer from quadratic complexity in self-attention mechanisms, imposing structural limitations. These challenges are often mitigated by custom implementations designed to reduce computational overhead or by introducing architectural modifications [21–24]. In addition to architecture customization, some studies restrict their models to a single modality, while others incorporate handcrafted features to further enhance performance [25]. There is also active research focused on hybrid models, which combine the strengths of both CNN and transformer architectures to leverage their complementary characteristics [26,27].

2.2. Foundation models

Foundation models are a class of models designed to be highly generalizable and scalable across a wide range of tasks and domains. They are characterized by their ability to be trained on massive amounts of data, often in an unsupervised or self-supervised manner, allowing them to learn versatile and reusable representations. These models, typically built using architectures like transformers [13,28], can be fine-tuned for specific downstream tasks with minimal task-specific data, making them invaluable in domains where labeled data is scarce. Foundation models serve as a base upon which task-specific models can be developed, reducing the need for extensive retraining and enabling more efficient use of resources across applications [29,30].

2.3. The emergence of self-supervised learning

SSL has transformed machine learning by allowing models to learn from vast amounts of unlabeled data. Instead of relying on manually annotated datasets, which are expensive and time-consuming to generate, SSL uses pretext tasks to uncover intrinsic patterns and structures. These tasks can involve predicting missing parts of an image, differentiating between augmented versions of the same input [31,32], or reconstructing masked images [10,12]. This flexibility makes SSL particularly attractive for medical imaging, where annotations are not only limited but also require domain-specific expertise.

Medical imaging, often hindered by a shortage of labeled data, is an ideal application for SSL. With SSL, models can learn meaningful representations of medical images without explicit labels, facilitating efficient learning and reducing dependency on costly annotations [33]. This capability has the potential to enhance clinical applications, such as disease detection and diagnosis support.

2.4. Challenges, limitations, and future directions

Despite the promise of SSL, several challenges remain, especially in medical imaging. One major issue is the difficulty of curating large datasets that satisfy SSL's requirements while adhering to patient privacy regulations [34,35]. Medical data is sensitive, and the process of anonymizing and sharing it across institutions can be complex and time-consuming, leading to fragmented or uneven datasets. This hampers the ability to train models that generalize effectively across diverse patient populations [36,37].

Another challenge is the variability in medical data. Different imaging modalities (e.g., magnetic resonance, computed tomography, X-ray) capture distinct aspects of human anatomy, and the inherent anatomical differences between patients add complexity [38–40]. As a result, SSL models must be highly adaptable, and capable of capturing diverse features without being overly specialized to any one modality or task.

Future research directions in SSL for medical imaging involve overcoming these data limitations by integrating multimodal datasets. Combining imaging data with other types of medical information, such as genomic or clinical data, could provide models with a more holistic understanding of disease processes, leading to better predictions and insights [41,42]. Additionally, improvements in model interpretability and explainability will be crucial for the safe deployment of these models in clinical settings. Ethical and regulatory frameworks must also evolve to support the broader use of SSL in healthcare.

2.5. Successes of foundation models in medical imaging

Foundation models, built using SSL techniques, have already achieved significant breakthroughs in medical imaging. These models, trained on large, diverse datasets, have demonstrated strong generalization capabilities across different imaging tasks, often outperforming traditional supervised models. SSL's flexibility allows these models to learn robust representations that benefit a variety of downstream tasks, including image classification, segmentation, and anomaly detection.

SSL-based models like DINO have shown the ability to capture intricate features in medical imaging datasets [15,43]. By processing large datasets, these models distill complex data into representations that can be used for diagnostic purposes, detecting subtle patterns and biomarkers often missed by human experts.

Among the various applications of foundation models in medical imaging, digital pathology employs image-based models as feature

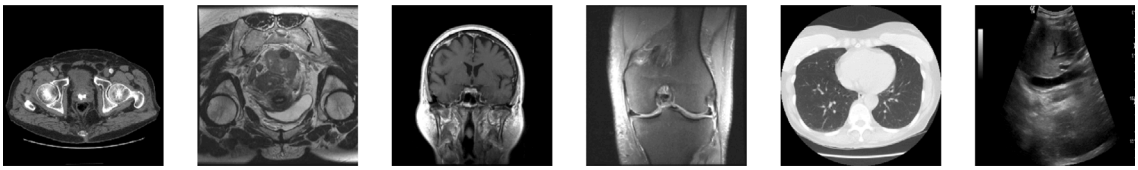

Fig. 1. Sample images extracted from the RadImageNet dataset.

extractors on individual patches from whole-slide images (WSIs) [44,45]. Direct training on WSIs is challenged by their extreme resolution and dimensionality, which impose prohibitive VRAM requirements unless specialized strategies are adopted. Transformer architectures with localized or dilated spatial attention render end-to-end WSI training computationally feasible [46]. Teacher–student frameworks reduce computational load during self-supervised pretraining by restricting the student network to partial or masked patch views [47]. Contrastive and masked pretraining methods dominate as self-supervised approaches in pathology, while supervised learning is generally reserved for fine-tuning due to inconsistent annotations and limited labeled samples across institutions.

The scalability of foundation models has important implications for AI-driven diagnostic tools, enabling the detection of subtle features in large datasets. As these models continue to evolve, they are expected to handle increasingly complex datasets and modalities, potentially paving the way for fully automated radiology workflows. However, further research is needed to ensure that these models generalize effectively across diverse clinical environments without compromising fairness or accuracy.

2.6. Multimodal foundation models in healthcare

In the context of healthcare, multimodal foundation models have emerged as a powerful tool for integrating various data sources such as medical images, clinical notes, genomic data, and laboratory results [48]. These models leverage the synergy between different modalities to provide a more comprehensive understanding of patient health, which can improve diagnosis and treatment planning [49]. By combining structured and unstructured data from multiple domains, multimodal foundation models are capable of capturing complex relationships between disparate sources of information, leading to better-informed clinical decision-making [50,51]. For example, models that integrate imaging data with genetic profiles can enhance the predictive accuracy for complex diseases such as cancer or neurodegenerative disorders [52]. Furthermore, foundation models can be applied to even more challenging tasks, such as 2D or 3D segmentation [53,54], where the challenges lie not only in learning effective representations but also in handling inherently high-dimensional data. This requires extreme care in terms of both performance and computational efficiency.

3. Materials and methods

In this section an overview of the materials and methods relative to this manuscript will be presented to provide a deeper description and explanation of the novelties proposed, serving as the foundation for the description of the experiments.

3.1. Datasets

This section will present the datasets employed for both the Radio DINO pretrain and downstream tasks.

RadImageNet

The dataset used for training our customized foundation model in radiomics is sourced from the extensive RadImageNet dataset [55]. This large-scale dataset comprises 1.35 million medical images spanning various imaging modalities, including computed tomography (CT), magnetic resonance (MR), and ultrasound imaging. These images cover eleven different anatomical regions, providing a comprehensive resource for the development of robust AI applications in radiologic imaging. The dataset contains 165 classes, encompassing a wide range of modalities and pathologies from the patients from whom the images were acquired. A notable challenge within the dataset is the significant class imbalance, a reflection of real-world medical scenarios where certain pathologies occur more frequently in the human body.

The RadImageNet dataset's annotations were meticulously curated by fellowship-trained and board-certified radiologists, ensuring high-quality and precise labels crucial for effective model training. This dataset's diverse and extensive nature enables the creation of models that can generalize well across multiple radiologic tasks.

In validation studies, RadImageNet pretrained models have demonstrated superior performance compared to those pretrained on ImageNet [56]. The extensive coverage of imaging modalities and anatomic regions, combined with expert annotations, positions the RadImageNet dataset as an exceptional foundation for training advanced models in radiomics. Visual examples of the RadImageNet dataset are illustrated in Fig. 1.

MedMNISTv2

The MedMNISTv2 dataset [57] is utilized as a benchmark for the downstream tasks to evaluate the effectiveness of the pretrained foundation model in specific medical image classification scenarios. MedMNISTv2 is a lightweight, open-source dataset designed for machine learning in the medical imaging domain, featuring 12 datasets for 2D and 6 datasets for 3D across different modalities and tasks, including X-rays, fundus images, dermoscopy, and so forth. These datasets are specifically tailored for tasks such as binary and multi-class classification, with labels curated by medical experts.

The MedMNISTv2 datasets span several important medical imaging tasks, such as pathology detection, organ segmentation, and disease classification. Each dataset is designed with different image sizes 28×28 up to 224×224, balancing computational efficiency with the ability to capture relevant medical features. In total, MedMNISTv2 encompasses over 700,000 images across a wide variety of clinical contexts, making it a valuable resource for validating the generalizability and performance of self-supervised models like Radio DINO.

For the downstream task evaluation, the MedMNISTv2 classification challenges are utilized to assess the capability of the pretrained model to transfer learned representations to specific clinical radiomics domains. Specifically, we selected the following datasets from the MedMNISTv2 collection: PneumoniaMNIST, BreastMNIST, OrganAMNIST, OrganCMNIST, and OrganSMNIST. These datasets span a range of radiologic tasks, such as pathology detection and organ classification, providing a diverse test bed for evaluating the model's generalization ability.

Visual examples of the selected datasets are illustrated in Fig. 2, while Table 1 presents the data distribution and class count for each dataset, along with detailed information about the different heterogeneous imaging data sources they utilize.

BUSI Dataset The BUSI dataset [58] is a publicly available collection of breast ultrasound images aimed at supporting research in breast cancer detection and classification. It consists of 780 images with a resolution of 500×500 pixels, collected from 600 female patients aged between 25 and 75 years. The dataset is divided into three categories: normal

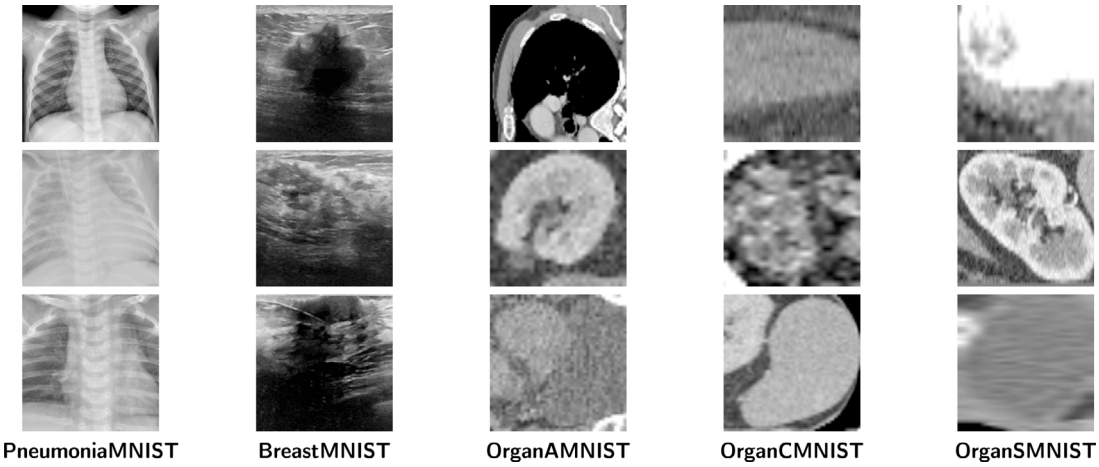

Fig. 2. Sample images extracted from the MedMNISTv2 dataset.

Table 1
Summary of data distribution and class count for each selected dataset, with details on modalities, tasks, number of classes, samples available, and splits provided(train/validation/test).

Dataset	Modality	Task	Classes	Total samples	Splits
PneumoniaMNIST	Chest X-ray	Binary-Class	2	5856	4708/524/624
BreastMNIST	Breast Ultrasound	Binary-Class	2	780	546/78/156
OrganAMNIST	Abdominal CT	Multi-Class	11	58,830	34,561/6491/17,778
OrganCMNIST	Abdominal CT	Multi-Class	11	23,583	12,975/2392/8216
OrganSMNIST	Abdominal CT	Multi-Class	11	25,211	13,932/2452/8827
RadImageNet	CT-MR-Ultrasound	Multi-Class	165	1,350,000	1,080,000/135,000/135,000
BUSI	Breast Ultrasound	Multi-Class	3	780	468/156/156

(133 images), benign (487 images), and malignant (210 images). Each image is accompanied by a ground truth segmentation mask, enabling its use for segmentation tasks in addition to classification. The images were collected using the LOGIQ E9 ultrasound system, and rigorous preprocessing steps were applied, including annotation validation by radiologists, to ensure data quality and consistency.

3.2. Vision Transformers (ViTs)

Transformers have revolutionized deep learning, especially in the field of natural language processing (NLP), by utilizing a self-attention mechanism that effectively captures long-range dependencies in sequential data [59]. This architecture replaces traditional recurrent and convolutional structures, allowing models to process entire sequences in parallel. The success of transformers in NLP inspired researchers to explore their potential in computer vision tasks, leading to the development of Vision Transformers.

At the core of transformers, including ViTs, is the multi-head self-attention mechanism. This mechanism works by computing attention scores between every pair of input tokens (or patches, in the case of ViTs) [59]. The attention score determines the relevance of one token to another, and multiple heads allow the model to capture different types of relationships in parallel. For each head, the input is projected into query, key, and value vectors. The attention is then computed as a weighted sum of the values, where the weights are derived from the dot product of queries and keys.

Mathematically, for a set of input embeddings X, the self-attention is computed as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

where Q, K, and V are the query, key, and value matrices derived from X, and d_k is the dimensionality of the key vectors [59].

Vision Transformers adapt the transformer architecture for image-based tasks by treating an image as a sequence of patches. Each image is divided into fixed-size non-overlapping patches, which are then flattened and linearly embedded into tokens. These tokens are analogous to words in NLP tasks. Positional embeddings are added to retain spatial information since transformers lack an inherent understanding of the positional structure of an image [60].

The embedded patches are then passed through a series of transformer encoder layers, where the multi-head self-attention mechanism is used to model relationships between patches. This approach allows ViTs to capture both local and global features of the image early in the processing pipeline. Unlike traditional convolutional neural networks (CNNs), which progressively build hierarchical features through localized convolutions, ViTs can model global dependencies from the start, leading to competitive performance in image classification tasks [60].

3.3. DINO and DINOv2 models

Self-supervised learning has gained significant traction in computer vision by enabling models to learn rich representations from unlabeled data. Among the successful methods in this domain, the DINO (Distillation with No Labels) and its successor, DINOv2, stand out as powerful frameworks that leverage the power of Vision Transformers for unsupervised learning [7].

DINO

DINO was introduced as a method for self-supervised learning using ViTs. The key idea behind DINO is to train a model in a self-distillation setup, where a "student" model learns to match the output of a "teacher" model, both operating on different views (augmentations) of the same image [7]. No labels are required in this process, and the system relies on augmentations of the input data to generate diverse views that the student and teacher must align.

In the DINO training process, the teacher model produces class-like outputs for multiple views of an image, which guide the student model during training. The student is updated to match the teacher's output, but the teacher model is updated more slowly through an exponential moving average (EMA) of the student's weights. As training progresses,

the student learns to produce consistent representations across these augmented views, enabling it to capture the semantics of the image [7]. DINO demonstrates that ViTs, even in a self-supervised setting, can discover semantic features such as object parts and boundaries without the need for any explicit supervision. This method significantly improved upon prior self-supervised techniques, especially in downstream tasks like image classification and object detection.

DINOv2

DINOv2 builds upon the success of DINO by addressing some of its limitations and further improving the quality of the learned representations [8]. DINOv2 introduces several key modifications to improve the robustness and scalability of the model. First, it uses improved architectural designs by utilizing more advanced ViT architectures and larger-scale training setups to extract better features from images. Second, the training pipelines have been optimized to handle larger datasets and more complex augmentations, leading to richer and more generalizable representations. Finally, DINOv2 refines the multi-crop strategy, which uses multiple smaller crops of the image during training to achieve better consistency between different views, resulting in superior feature alignment across diverse augmentations [8].

3.4. Classification metrics

The performance of the classification experiments was evaluated using five key metrics: Accuracy, Precision, Recall, F1 Score, and Area Under the Receiver Operating Characteristic Curve (AUC). These metrics provide a comprehensive view of the model's ability to correctly classify instances, balancing both correctness and robustness in predictions.

When comparing classification results to true target values, we can determine whether an instance belongs to one of the following categories:

- **True Negatives (TN)** is the number of instances from the negative class that have been correctly predicted as negative.
- **False Positives (FP)** is the number of instances from the negative class that has been erroneously predicted as positive.
- **False Negatives (FN)** is the number of instances from the positive class that has been erroneously predicted as negative.
- **True Positives (TP)** is the number of instances from the positive class that have been correctly predicted as positive.

Accuracy is the ratio of correctly predicted instances to the total number of instances in the dataset. While it provides a general measure of performance, it can be misleading for imbalanced datasets, where the majority class may dominate the predictions. Accuracy is defined as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision measures the proportion of true positive predictions out of all positive predictions made by the model. This metric is crucial in scenarios where false positives are costly. It is computed as:

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall, also known as sensitivity, is the ratio of correctly predicted positive instances to all actual positive instances. It reflects the model's ability to capture all relevant cases, which is essential in scenarios where missing positive cases is costly. Recall is calculated as:

$$\text{Recall} = \frac{TP}{TP + FN}$$

The F1 Score (F1) is the harmonic mean of Precision and Recall, providing a balanced metric that accounts for both false positives and false negatives. It is particularly useful in cases of class imbalance. The F1 score is given by:

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Finally, the AUC (Area Under the ROC Curve) represents the area under the Receiver Operating Characteristic curve, which plots the true positive rate against the false positive rate at various threshold levels. AUC measures the model's ability to distinguish between positive and negative classes, with higher values indicating better discriminatory power. An AUC closer to 1.0 signifies strong classification performance.

3.5. Segmentation metrics

For segmentation tasks, the model's performance is typically evaluated using the following metrics: Intersection over Union (IoU) and Dice Similarity Coefficient (Dice). These metrics measure the overlap between the predicted and ground truth segmentation masks.

Intersection over Union is a metric used to evaluate the accuracy of an object detector on a particular dataset. It measures the overlap between the predicted segmentation mask and the ground truth mask. The IoU is computed as:

$$\text{IoU} = \frac{|A \cap B|}{|A \cup B|}$$

where A is the predicted mask and B is the ground truth mask. A higher IoU indicates better performance, with 1.0 representing perfect overlap.

Dice Similarity Coefficient is another metric for measuring the similarity between two sets. It is especially popular in image segmentation tasks as it gives more weight to smaller regions compared to IoU. The Dice coefficient is defined as:

$$\text{Dice} = \frac{2 \times |A \cap B|}{|A| + |B|}$$

where A is the predicted mask and B is the ground truth mask. Like IoU, a higher Dice score indicates better performance, with a perfect match yielding a value of 1.0.

4. Experimental setup

The experiments were conducted using a GPU on-demand service, employing two NVIDIA A100 GPUs for training each equipped with 80 gb VRAM. This setup allows for efficient handling of large models and complex architectures, ensuring scalable performance. For the experiments, we utilized the official repositories for DINO[1] and DINOv2,[2] available on GitHub.

Both Radio DINO and Radio DINOV2 were trained across all their size variants using a learning rate of 3×10^{-4} for 200 epochs. Due to memory constraints, the Radio DINOV2 and Radio DINO base models were trained with a batch size of 128, while the Radio DINO tiny and small models were trained with batch sizes of 512 and 256, respectively. The key distinction between Radio DINO and Radio DINOV2 is the patch size: Radio DINOV2 was trained with a patch size of 14, whereas Radio DINO used a patch size of 16 to replicate the experimental settings proposed by the respective authors. For DINO we also investigate the impact of patch size by training the tiny and small models with a patch size of 8, dividing by 3 the batch size to satisfy the memory requirements, we denote such versions as Radio DINO tiny sp and Radio DINO small sp. We provide a brief description of each model's characteristics in Table 2, and a schematic representation of the entire proposed pipeline is shown in Fig. 3.

Our pipeline consists of two main phases: the pretraining phase and the evaluation phase.

In the pretraining phase, we tune the parameters of the Vision Transformer by attaching either a DINO or DINOv2 head, depending on the chosen pretraining strategy. The model iterates through images from the RadImageNet dataset, extracting meaningful representations without relying on any labeled data.

[1] https://github.com/facebookresearch/dino.
[2] https://github.com/facebookresearch/dinov2.

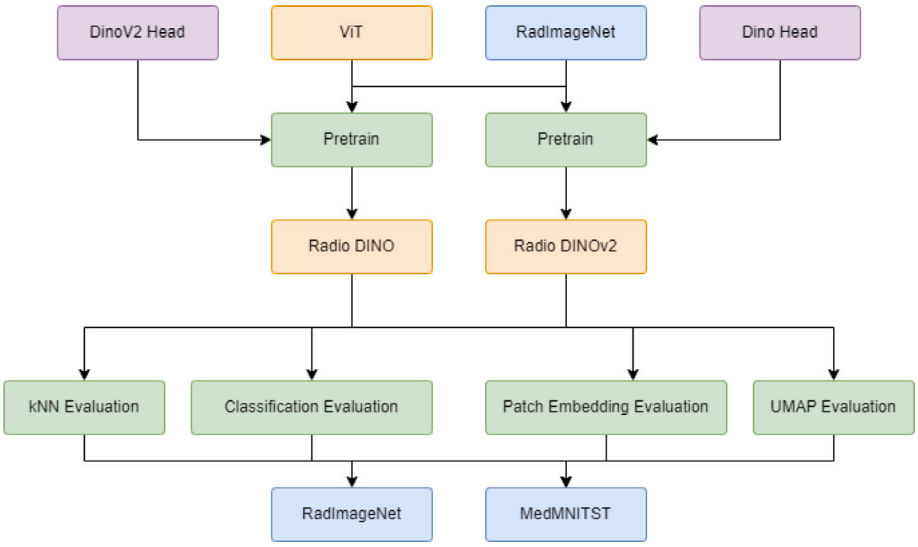

Fig. 3. Pipeline schema illustrating different semantic concepts. Purple blocks represent Self Supervised Learning heads, yellow blocks denote deep learning models, green blocks indicate the various tasks, and blue blocks correspond to the datasets.

In the evaluation phase, we leverage the rich semantic embeddings from the tuned models by removing the pretraining heads. For both Radio DINO and Radio DINOv2, we first perform a kNN evaluation to determine which model offers the best representation and compare these representations against other state-of-the-art pretrained models. We also visualize the PCA of the embeddings to assess whether the structures of the test images are well-represented in the RGB space.

Finally, we fine-tune our Radio DINO models on downstream tasks using the MedMNIST dataset to demonstrate the models' capacity to generalize. This fine-tuning process highlights how effectively the pre-trained models' representations serve as a strong foundation for further learning and task-specific performance.

4.1. Classification setup

We applied standard preprocessing to the input images. During training, images were resized to 224×224 pixels and augmented using the AutoAugment strategy. All images were normalized using ImageNet mean and standard deviation. For validation and testing, the same resizing and normalization were used without augmentation to ensure unbiased evaluation.

Training involved two phases: a warmup phase and the main training phase. The warmup used a reduced learning rate (5×10^{-6}) for the first five epochs. In the main phase, the learning rate was restored to 1×10^{-4}, with a cosine annealing scheduler applied to progressively reduce the rate. The AdamW optimizer, with a weight decay of 0.01, was used, and CrossEntropy loss guided both phases.

Training spanned 30 epochs, with early stopping triggered if validation loss failed to improve for 10 epochs. A batch size of 256 was used, with automatic mixed precision (AMP) to speed up training and reduce memory usage. Gradient clipping, capped at a maximum norm of 1, was applied for stability.

Experiments were repeated using a 5-fold classification strategy.

5. Evaluation

We conducted a comprehensive evaluation of our Radio Dino models, comparing their performance with several state-of-the-art architectures, including the original Dino and DinoV2 models, supervised and self-supervised models in feature evaluation and downstream tasks.

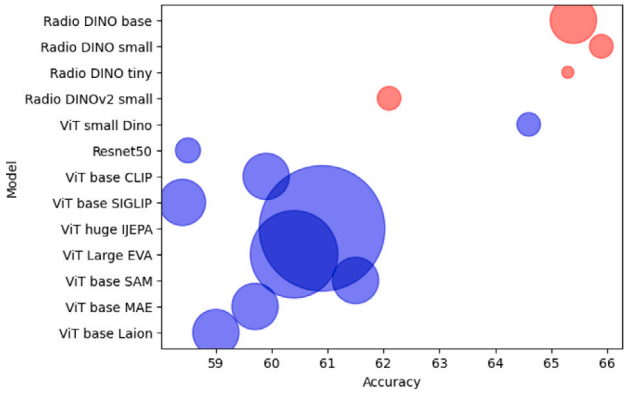

Fig. 4. Scatter plot of KNN evaluation with 1 neighbor on the RadImageNet dataset [55]. The area of each marker is proportional to the number of parameters in the corresponding model.

Table 2
Embedding sizes and patch sizes for different Radio DINO and Radio DINOV2 variants.

Model variant	Embedding size	Patch size	Parameters(M)
Radio DINOv2	384	14	22.2
Radio DINO tiny	192	16	5.8
Radio DINO tiny sp	192	8	5.9
Radio DINO small	384	16	22.2
Radio DINO small sp	384	8	22.3
Radio DINO base	768	16	86

5.1. Feature evaluation

We decided to conduct our feature evaluation experiments using k-nearest neighbors (kNN) classification to gauge the effectiveness of the extracted features. By performing kNN classification, we were able to analyze how well our model clusters similar samples within the feature space. The full set of kNN experiments, covering various k parameter configurations, is detailed in Table 3, where the best results are highlighted in bold. Finally, to offer a more visual representation of our model's capabilities, Fig. 4 showcases the accuracy of different comparison models versus our proposed foundation models, specifically with the k parameter set to 1.

Table 3
Performance metrics for analyzed models across different k configurations on the RadImageNet dataset [55].

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)
k = 1				
[13] ViT base CLIP ImageNet finetuned	59.04	27.15	27.18	26.76
[12] ViT base MAE	59.72	30.28	30.13	29.69
[61] ViT base SAM	61.56	30.70	30.01	29.92
[11] ViT large EVA	60.43	30.37	29.78	29.64
[10] ViT huge IJEPA	60.93	30.82	30.86	30.34
[62] ViT base SIGLIP	58.43	27.33	27.12	26.85
[13] ViT base CLIP	59.95	29.08	28.79	28.48
[8] ViT small DINOv2	61.59	30.24	29.62	29.52
[63] ResNet50 ImageNet finetuned	58.55	26.79	26.77	26.45
[7] ViT small DINO	64.62	34.89	33.98	33.82
(Our) Radio DINO base	65.50	35.10	35.80	34.80
(Our) Radio DINO small	65.90	**35.90**	**36.20**	**35.40**
(Our) Radio DINO small sp	62.90	32.00	32.40	31.80
(Our) Radio DINO tiny	**66.20**	35.80	36.00	35.30
(Our) Radio DINO tiny sp	65.30	33.40	33.60	33.10
k = 5				
[13] ViT base CLIP ImageNet finetuned	60.43	26.40	22.02	22.76
[12] ViT base MAE	60.78	29.68	24.78	25.60
[61] ViT base SAM	62.80	30.78	25.06	25.97
[11] ViT large EVA	61.45	30.75	24.28	25.19
[10] ViT huge IJEPA	61.83	30.64	25.70	26.59
[62] ViT base SIGLIP	59.84	27.36	22.55	23.33
[13] ViT base CLIP	61.12	29.09	23.40	24.33
[8] ViT small DINOv2	62.89	31.31	24.58	25.64
[63] ResNet50 ImageNet finetuned	59.94	26.25	21.37	22.06
[7] ViT small DINO	65.85	34.21	28.74	30.02
(Our) Radio DINO base	66.50	34.20	30.30	30.90
(Our) Radio DINO small	66.80	**35.90**	**30.80**	**31.60**
(Our) Radio DINO small sp	63.90	32.20	27.20	28.00
(Our) Radio DINO tiny	**67.10**	35.30	30.30	31.20
(Our) Radio DINO tiny sp	66.40	33.00	28.20	29.00
k = 20				
[13] ViT base CLIP ImageNet finetuned	62.41	31.38	20.57	21.28
[12] ViT base MAE	62.52	32.48	23.11	23.90
[61] ViT base SAM	64.99	35.22	23.52	24.59
[11] ViT large EVA	63.02	32.31	22.27	23.19
[10] ViT huge IJEPA	63.61	34.05	23.96	24.98
[62] ViT base SIGLIP	61.72	30.08	20.86	21.57
[13] ViT base CLIP	63.22	33.24	21.82	22.77
[8] ViT small DINOv2	64.89	34.98	22.65	23.67
[63] ResNet50 ImageNet finetuned	62.30	31.09	20.32	20.96
[7] ViT small DINO	67.67	39.61	26.41	27.79
(Our) Radio DINO base	68.50	39.20	28.60	29.60
(Our) Radio DINO small	68.80	**40.10**	**29.00**	**30.00**
(Our) Radio DINO small sp	65.90	36.50	25.60	26.80
(Our) Radio DINO tiny	**68.90**	39.10	28.30	29.60
(Our) Radio DINO tiny sp	68.30	38.50	26.90	28.00

The results demonstrate a clear superiority of our Radio DINO and Radio DINOv2 models over other off-the-shelf models. However, it is noteworthy that the standard DINO features outperformed our Radio DINOv2 features. This observation led us to focus our research and computational efforts on the Radio DINO series, underscoring the importance of incorporating register mechanisms in vision transformers [64].

5.2. Classification results

In this section, we present and discuss the classification performance of our Radio DINO model. When evaluated across the selected datasets, Radio DINO outperformed the previous state-of-the-art solutions in 4 out of 5 datasets, while delivering comparable results on the remaining dataset. We show by One-Sample t-Test the relevance of our results by using as reference value the best results form previous works.

For the BreastMNIST dataset, Radio DINO exceeded the previous best model, Med ViT small [65], by achieving an improvement of 1.97% in accuracy (p-value of 0.0029). A complete breakdown of the results for this dataset can be found in Table 4.

On the OrganAMNIST, OrganCMNIST, and OrganSMNIST datasets, Radio DINO achieved an approximate 3% improvement across all reference metrics compared to the former state-of-the-art [65]. A schematization of the results for this datasets can be found in Tables 6–8. Specifically the improvements for all three datasets in terms of accuracy were respectively: 2.1% p-value 0.0009, 2.89% p-value 0.0008, and 1.51% p-value 0.02

However, the PneumoniaMNIST dataset posed the greatest challenge for our model. Despite this, Radio DINO managed to deliver competitive results, with only a marginal drop of 0.57% in AUC compared to the previous best solution. Results are depicted in Table 5

Throughout the experiments, Radio DINO consistently exhibited low variance across the results. Interestingly, we observed that variance increased when larger models were employed. However, this increase is not an inherent limitation of Radio DINO but rather a well-known issue in deep learning models, as discussed in [66].

Table 4
Performance comparison of different models on the BreastMNIST dataset. The table reports accuracy, precision, recall, F1, and AUC, with an indication of the standard deviation when available. The best results are emphasized in bold.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)	AUC (%)
Resnet18 [57]	83.3	–	–	–	89.1
Resnet50 [57]	84.2	–	–	–	86.6
auto-sklearn [57]	80.3	–	–	–	83.6
AutoKeras [57]	83.1	–	–	–	87.1
Google AutoML Vision [57]	86.1	–	–	–	91.9
DenseNet121 [67]	86.3	–	–	–	90.1
Swin Transformer [68]	86.92 ± 1.04	–	–	–	86.26 ± 1.38
R-LLM [69]	87.17	–	–	–	88.23
Med ViT tiny [65]	89.6	–	–	–	93.4
Med ViT small [65]	89.7	–	–	–	93.8
Med ViT base [65]	88.3	–	–	–	92.9
DINO small	74.57 ± 4.17	66.20 ± 10.37	58.79 ± 5.86	59.14 ± 7.56	71.78 ± 4.41
DINO base	83.12 ± 0.98	81.49 ± 1.3	72.91 ± 1.79	75.48 ± 1.77	83.16 ± 1.05
DINOv2 small	73.50 ± 2.67	62.14 ± 15.28	53.55 ± 4.12	50.63 ± 6.57	67.28 ± 7.19
DINOv2 base	69.02 ± 7.59	50.49 ± 14.08	51.98 ± 3.22	49.07 ± 6.35	63.22 ± 13.01
Radio DINO tiny sp	87.18 ± 0.37	85.44 ± 0.36	80.7 ± 0.69	82.6 ± 0.59	89.97 ± 1.66
Radio DINO small sp	84.62 ± 0.98	83.98 ± 2.97	79.7 ± 3.29	80.15 ± 1.97	88.37 ± 1.62
Radio DINO tiny	90.38 ± 0.81	89.04 ± 1.29	85.9 ± 1.27	87.29 ± 1.1	92.31 ± 0.98
Radio DINO small	**91.67 ± 0.68**	90.81 ± 0.69	**87.53 ± 1.21**	**88.98 ± 1.0**	**95.55 ± 1.55**
Radio DINO base	90.38 ± 1.84	**91.11 ± 3.15**	87.41 ± 4.93	87.69 ± 3.5	95.51 ± 1.7

Table 5
Performance comparison of different models on the PneumoniaMNIST dataset. The table reports accuracy, precision, recall, F1, and AUC, with an indication of the standard deviation when available. The best results are emphasized in bold.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)	AUC (%)
Resnet18 [57]	86.4	–	–	–	95.6
Resnet50 [57]	88.4	–	–	–	96.2
auto-sklearn [57]	85.5	–	–	–	94.2
AutoKeras [57]	87.8	–	–	–	94.7
Google AutoML Vision [57]	94.6	–	–	–	99.1
GCNN-EC [67]	87.7	–	–	–	95.4
Swin Transformer [68]	91.54 ± 0.48	–	–	–	98.22 ± 0.37
R-LLM [69]	93.91	–	–	–	98.01
Med ViT tiny [65]	94.9	–	–	–	99.3
Med ViT small [65]	**96.1**	–	–	–	**99.5**
Med ViT base [65]	92.1	–	–	–	99.1
DINO small	83.71 ± 1.44	86.86 ± 0.17	79.22 ± 2.21	80.95 ± 2.14	93.89 ± 0.79
DINO base	90.70 ± 2.00	93.35 ± 1.25	87.69 ± 2.66	89.46 ± 2.41	98.52 ± 1.13
DINOv2 small	81.04 ± 1.73	83.63 ± 2.38	76.22 ± 2.12	77.75 ± 2.21	92.21 ± 1.19
DINOv2 base	79.49 ± 0.48	82.91 ± 0.51	73.93 ± 0.60	75.41 ± 0.66	91.72 ± 0.13
Radio DINO tiny sp	91.83 ± 1.53	93.69 ± 0.74	89.36 ± 2.14	90.88 ± 1.88	97.76 ± 0.68
Radio DINO small sp	89.42 ± 1.15	91.94 ± 0.51	86.24 ± 1.62	88.01 ± 1.48	97.55 ± 0.59
Radio DINO tiny	91.19 ± 1.11	93.63 ± 0.6	88.33 ± 1.5	90.07 ± 1.37	98.51 ± 0.2
Radio DINO small	91.83 ± 1.03	93.86 ± 0.54	89.27 ± 1.42	90.86 ± 1.27	98.86 ± 0.23
Radio DINO base	93.91 ± 1.11	**95.25** ± 0.66	**92.05 ± 1.51**	**93.29 ± 1.31**	98.93 ± 0.45

Table 6
Performance comparison of different models on the OrganAMNIST dataset. The table reports accuracy, precision, recall, F1, and AUC, with an indication of the standard deviation when available. The best results are emphasized in bold.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)	AUC (%)
Resnet18 [57]	95.1	–	–	–	99.8
Resnet50 [57]	94.7	–	–	–	99.8
auto-sklearn [57]	76.2	–	–	–	96.3
AutoKeras [57]	90.5	–	–	–	99.4
Google AutoML Vision [57]	88.6	–	–	–	99.0
DenseNet121 [67]	93.5	–	–	–	99.7
R-LLM [69]	95.22	–	–	–	99.78
Med ViT tiny [65]	93.1	–	–	–	99.5
Med ViT small [65]	92.8	–	–	–	99.6
Med ViT base [65]	94.3	–	–	–	99.7
DINO small	96.45 ± 0.58	96.10 ± 0.64	95.87 ± 0.53	95.93 ± 0.60	99.88 ± 0.03
DINO base	96.79 ± 0.44	96.64 ± 0.37	96.24 ± 0.79	96.39 ± 0.61	99.87 ± 0.04
DINOv2 small	95.07 ± 0.42	94.79 ± 0.53	94.58 ± 0.40	94.55 ± 0.41	99.83 ± 0.01
DINOv2 base	94.56 ± 1.13	94.36 ± 1.04	94.13 ± 1.21	94.19 ± 1.13	99.76 ± 0.08
Radio DINO tiny sp	95.74 ± 0.89	95.53 ± 0.94	95.14 ± 1.01	95.28 ± 1.02	99.85 ± 0.09
Radio DINO small sp	96.2 ± 0.46	96.14 ± 0.66	95.89 ± 0.68	95.97 ± 0.67	99.88 ± 0.06
Radio DINO tiny	96.4 ± 0.45	96.21 ± 0.52	95.96 ± 0.64	96.03 ± 0.59	99.9 ± 0.03
Radio DINO small	96.83 ± 0.37	96.71 ± 0.4	96.39 ± 0.33	96.47 ± 0.36	99.92 ± 0.03
Radio DINO base	**97.35 ± 0.55**	**97.22 ± 0.55**	**97.22 ± 0.64**	**97.2 ± 0.61**	**99.93 ± 0.03**

Table 7
Performance comparison of different models on the OrganCMNIST dataset. The table reports accuracy, precision, recall, F1, and AUC, with an indication of the standard deviation when available. The best results are emphasized in bold.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)	AUC (%)
Resnet18 [57]	92.0	–	–	–	99.4
Resnet50 [57]	91.1	–	–	–	99.3
auto-sklearn [57]	82.9	–	–	–	97.6
AutoKeras [57]	87.9	–	–	–	99.0
Google AutoML Vision [57]	87.7	–	–	–	98.8
EfficientNetB0 [67]	90.5	–	–	–	99.2
Med ViT tiny [65]	90.1	–	–	–	99.1
Med ViT small [65]	91.6	–	–	–	99.3
Med ViT base [65]	92.2	–	–	–	99.4
DINO small	93.67 ± 0.26	93.61 ± 0.41	92.44 ± 0.35	92.91 ± 0.36	99.67 ± 0.16
DINO base	94.28 ± 1.44	94.19 ± 1.82	94.13 ± 1.23	94.06 ± 1.62	99.72 ± 0.05
DINOv2 small	91.37 ± 0.49	90.55 ± 0.77	89.39 ± 0.67	89.83 ± 0.72	99.38 ± 0.28
DINOv2 base	85.73 ± 1.36	83.68 ± 2.17	82.11 ± 2.12	82.61 ± 2.20	98.84 ± 0.21
Radio DINO tiny sp	93.28 ± 0.1	93.36 ± 0.08	92.4 ± 0.18	92.78 ± 0.13	99.75 ± 0.02
Radio DINO small sp	93.73 ± 0.73	93.6 ± 0.5	92.86 ± 1.05	93.11 ± 0.8	99.77 ± 0.05
Radio DINO tiny	93.99 ± 0.45	93.96 ± 0.37	93.02 ± 0.55	93.37 ± 0.46	99.79 ± 0.03
Radio DINO small	94.3 ± 0.31	93.96 ± 0.18	93.69 ± 0.4	93.63 ± 0.25	99.8 ± 0.02
Radio DINO base	**95.11 ± 0.71**	**94.91 ± 0.75**	**94.49 ± 0.83**	**94.57 ± 0.82**	**99.86 ± 0.04**

Table 8
Performance comparison of different models on the OrganSMNIST dataset. The table reports accuracy, precision, recall, F1, and AUC, with an indication of the standard deviation when available. The best results are emphasized in bold.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)	AUC (%)
Resnet18 [57]	77.8	–	–	–	97.4
Resnet50 [57]	78.5	–	–	–	97.5
auto-sklearn [57]	67.2	–	–	–	94.5
AutoKeras [57]	81.3	–	–	–	97.4
Google AutoML Vision [57]	74.9	–	–	–	96.4
Resnet18 [67]	81.3	–	–	–	97.4
Med ViT tiny [65]	78.9	–	–	–	97.2
Med ViT small [65]	80.5	–	–	–	**98.7**
Med ViT base [65]	80.6	–	–	–	97.3
DINO small	80.11 ± 1.51	76.12 ± 2.12	74.49 ± 1.27	74.66 ± 1.35	97.97 ± 0.26
DINO base	82.44 ± 0.44	78.54 ± 0.31	77.15 ± 0.67	77.59 ± 0.27	97.21 ± 0.08
DINOv2 small	77.87 ± 0.62	72.63 ± 0.73	72.16 ± 0.67	71.82 ± 0.80	97.62 ± 0.12
DINOv2 base	73.97 ± 1.66	68.20 ± 1.96	67.25 ± 2.18	66.70 ± 2.17	97.01 ± 0.31
Radio DINO tiny sp	80.57 ± 0.68	76.71 ± 0.78	75.87 ± 0.89	76.1 ± 0.91	97.68 ± 0.15
Radio DINO small sp	79.98 ± 0.45	75.8 ± 0.07	74.83 ± 0.83	74.97 ± 0.38	97.55 ± 0.11
Radio DINO tiny	81.53 ± 0.61	78.1 ± 0.53	77.43 ± 1.04	76.88 ± 0.64	98.06 ± 0.09
Radio DINO small	82.3 ± 0.81	78.63 ± 0.91	77.67 ± 0.84	77.73 ± 1.2	98.27 ± 0.12
Radio DINO base	**82.81 ± 0.96**	**79.51 ± 1.14**	**77.75 ± 1.1**	**78.15 ± 1.17**	98.28 ± 0.15

6. Feature visualization and explainability

In this section, we present a detailed analysis of the features extracted from our Radio DINO models. The primary goal is to assess whether these features can be directly utilized without fine-tuning by examining the distribution of data across different datasets within the feature space. Additionally, we provide insights from the fine-tuned models to evaluate the ability of Radio DINO to accurately identify key characteristics and patterns, enabling the correct classification of radiological images into their respective intrinsic categories.

6.1. Embedding PCA visualization

To visualize the learned feature space, we applied Principal Component Analysis (PCA) to the embeddings generated by Radio DINO. This method projects high-dimensional embeddings onto a 2D plane, offering insights into the clustering and separation of different classes within the feature space.

As shown in Fig. 5, the embeddings from our model demonstrate strong utility and semantic coherence. The figure depicts a visualization of three principal components of the patch embeddings, represented as color channels. These embeddings reveal a significant semantic differentiation between regions of the analyzed area. Despite the model not being explicitly trained to distinguish such regions, a clear separation between the eyes, brain, and skull is evident.

From Fig. 5, several characteristics can be observed across all models, except for the randomly initialized one, which does not exhibit a distinct brain region patch cluster. The other models, however, not only differentiate between the brain and skull regions but also show a clear separation of the brainstem. These emerging capabilities are further emphasized in the patch embeddings from Radio DINOV2, which employs a smaller patch size among the selected models, enabling a more detailed representation.

Although Radio DINOV2 provides a more refined visualization, the embeddings also show high-norm tokens forming a square-like structure. This phenomenon could be attributed to the model's need to store information in selected patches of the image, as discussed in [9].

6.2. UMAP visualization

In addition to PCA, we applied Uniform Manifold Approximation and Projection (UMAP) for dimensionality reduction and visualization. UMAP is often capable of providing a more faithful representation of high-dimensional structures in a lower-dimensional space, potentially revealing more nuanced relationships between samples.

Fig. 6 illustrates the UMAP visualization of embeddings from our proposed foundation model on the RadImageNet test set, showing a distinct separation of class subgroups based on their respective regions and pathologies. Despite the high performance demonstrated

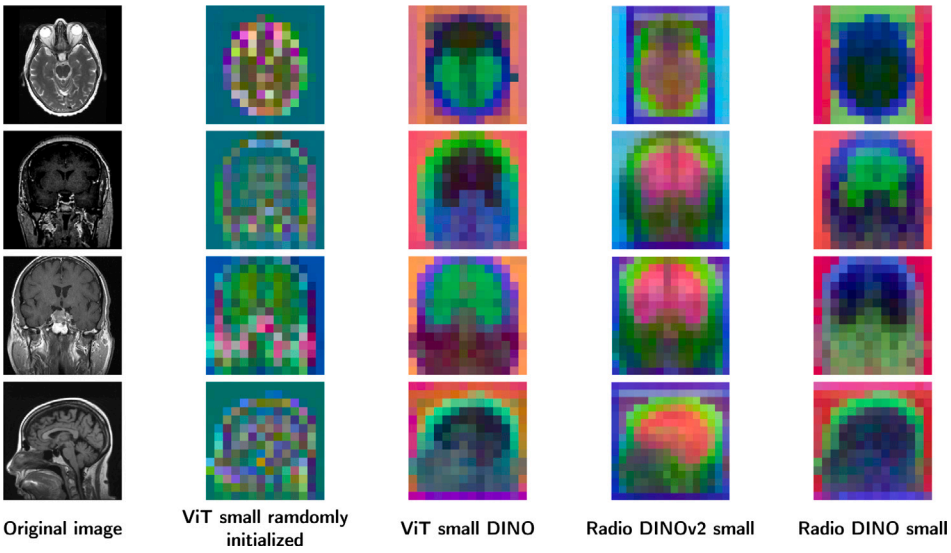

Fig. 5. PCA visualizations of patch embeddings for four different models across four radiological images from the RadImageNet dataset. The first column shows the original images, while the following columns display PCA visualizations of the embeddings produced by a randomly initialized ViT model, ViT small DINO, Radio DINOv2 small, and Radio DINO small models, respectively.

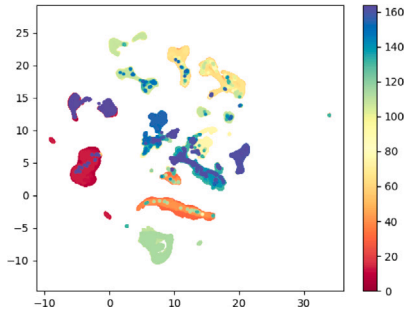

Fig. 6. UMAP visualization of Radio DINO small on the RadImageNet test set. The classes are depicted as numbers sorted by name.

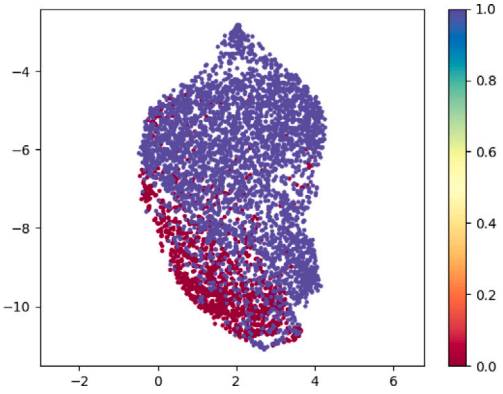

Fig. 7. UMAP visualization of Radio DINO small on the PneumoniaMNIST test set. The classes are depicted as numbers sorted by name. The normal class is represented by the color purple, while the pneumonia class is represented by red.

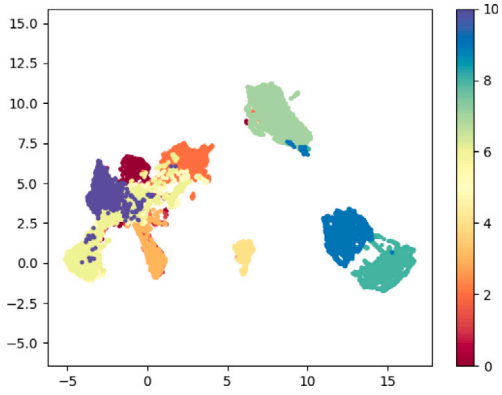

Fig. 8. UMAP visualization of Radio DINO small on the OrganAMNIST test set. The classes are depicted as numbers sorted by name.

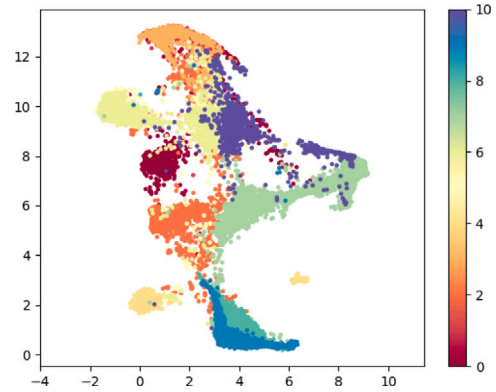

Fig. 9. UMAP visualization of Radio DINO small on the OrganCMNIST test set. The classes are depicted as numbers sorted by name.

in Table 3, similar classes remain in close proximity within the embedding space. This overlap might be mitigated by increasing the embedding dimensionality, as class-wise embedding collapse may necessitate higher-dimensional features for precise separation.

Moreover, in Figs. 7–9, we present UMAP visualizations for the PneumoniaMNIST, OrganAMNIST, and OrganCMNIST datasets. Notably, even for these datasets, our Radio DINO model successfully separates images from different datasets without explicit training to do so this is purely a result of the pretraining process. This capability is especially evident in Fig. 7, where two distinct clusters naturally emerge.

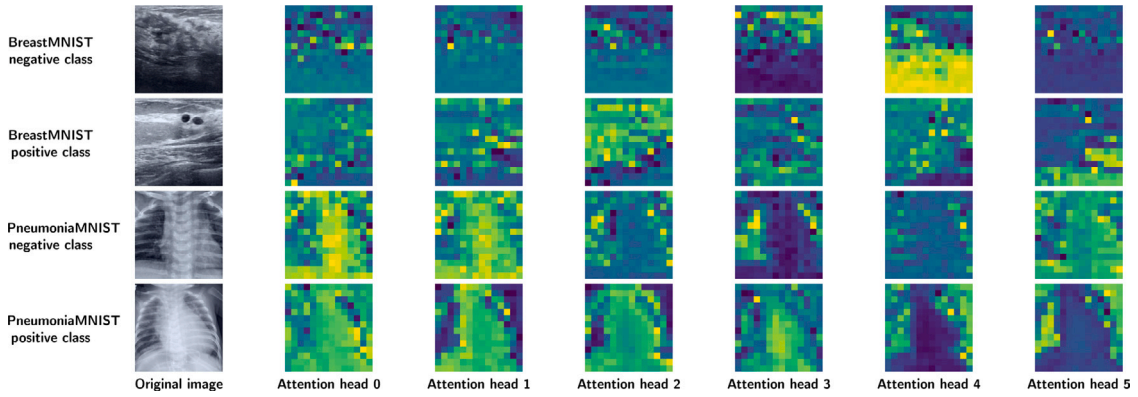

Fig. 10. Finetuned Radio DINO attention maps visualization.

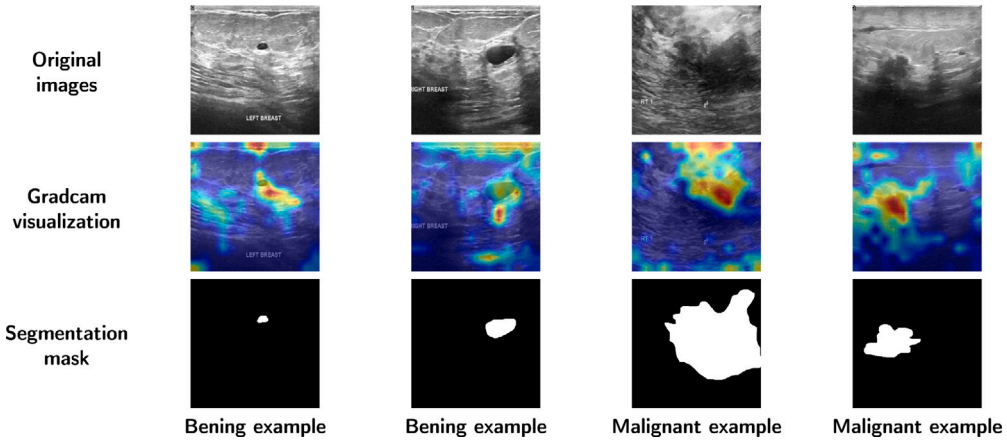

Fig. 11. Finetuned Radio DINO Grad-CAM visualization on the BUSI dataset.

6.3. Attention maps visualization

In this section, we present attention map visualizations from the heads of the final layer of the Radio DINO small model, fine-tuned on selected classification datasets from the MedMNISTv2 collection. The self-attention maps are shown for the [CLS] token query.

As illustrated in Fig. 10, the model focuses on nodule-like structures in the BreastMNIST dataset, which are key indicators of malignancies in breast ultrasound images. The visualization highlights patches near these round structures, with high norm values emphasizing the importance of these regions for accurate classification.

For the PneumoniaMNIST dataset, the model attends not only to the consolidation areas solid, opaque regions that appear white on lung images but also to air bronchograms, which are clearly visible in attention heads 4 and 5 for a pneumonia-positive patient, as shown in Fig. 10. These attention maps highlight the model's ability to focus on diagnostically relevant areas in the images.

7. Proposed evaluation methodology

We propose a comprehensive evaluation methodology to rigorously assess the performance of our foundation model against state-of-the-art baselines. This methodology evaluates not only classification and segmentation performance but also the generalizability and robustness of learned features across varying image scales. The BUSI dataset [58] was selected for its multi-class annotations and segmentation masks, providing a strong foundation for thorough comparative analysis. Our models are fine-tuned on the BUSI dataset following the experimental protocols used with MedMNISTv2 [57]. From the trained models, we calculate Grad-CAM visualizations for each image. These are quantitatively compared to the corresponding ground truth segmentation masks using the Dice coefficient, offering an objective measure of the models' localization capabilities. In addition to Grad-CAM analysis, we evaluate the generalizability of the learned features by training a lightweight segmentation head on frozen model features. This approach minimizes the number of trainable parameters, ensuring that any performance differences reflect the quality of the backbone representations. The segmentation head excludes the ViT-Adapter [70], isolating the intrinsic generalizability of the evaluated models without additional architectural enhancements. Considering the variability in clinical imaging resolutions, we also assess model robustness across different image scales. Features are extracted from the CLS token of the trained models using input resolutions of 448×448, 224×224, 112×112, and 56×56. A logistic regression classifier is trained on these features to evaluate classification performance, providing insights into the scale-invariance of the learned representations.

7.1. Fine-tuning performance evaluation

To evaluate the performance of our models under supervised fine-tuning, we conducted experiments on the BUSI dataset, measuring Accuracy, Precision, Recall, F1 score, and AUC. Fine-tuning was performed on multiple variants of the DINO and Radio DINO models. As shown in Table 9, the Radio DINO base model consistently outperforms other variants, achieving the highest accuracy of 92.41%, an F1 score of 91.73%, and an AUC of 98.35%. This strong performance reflects the model's robust feature representation and effective learning during fine-tuning.

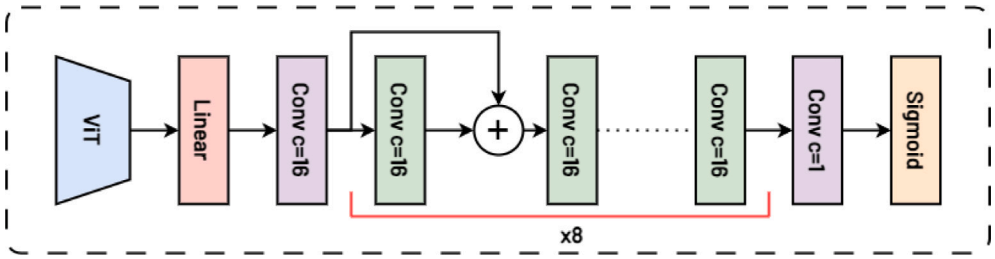

Fig. 12. Architecture of the proposed segmentation head. Convolutional layers highlighted in green are repeated 8 times. The parameter c denotes the number of output channels. The in-depth description of the implementation details is depicted in Section 7.2.

Table 9
Performance metrics of different models after supervised fine-tuning on the BUSI dataset.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)	AUC (%)
DINO small	53.16	36.86	41.58	36.60	55.63
DINO base	88.61	88.48	85.00	86.38	97.28
Radio DINO tiny sp	81.01	88.42	75.00	78.63	93.90
Radio DINO small sp	86.08	91.11	81.67	84.77	96.97
Radio DINO tiny	75.95	84.24	69.15	71.51	91.49
Radio DINO small	83.54	89.78	78.33	81.83	94.85
Radio DINO base	**92.41**	**94.25**	**90.00**	**91.73**	**98.35**

Table 10
Segmentation performance comparison on the BUSI dataset based on Dice score, IoU, and the number of trainable parameters.

Model	Dice (%)	IoU (%)	Trainable parameters (M)
DINO small	67.80	51.10	0.72
DINO base	65.80	49.60	0.80
Radio DINO tiny	68.90	52.80	**0.65**
Radio DINO tiny sp	66.40	50.10	**0.65**
Radio DINO small	68.30	52.40	0.72
Radio DINO small sp	**71.70**	**56.50**	0.70
Radio DINO base	69.50	53.50	0.80

7.2. Feature evaluation through segmentation

To expand our segmentation analysis, we propose a novel approach for segmenting images directly using frozen backbone features from the Radio DINO models. This minimalist method focuses on reducing the number of learnable parameters to ensure a fair comparison across models. The segmentation framework begins with a linear projection layer to adjust the extracted features' dimensionality to match the flattened patch size of the Vision Transformer architecture. For Radio DINO models with an 8-pixel patch size, the linear layer outputs 64-dimensional features, while for a 16-pixel patch size, it outputs 256-dimensional features. These features are then processed by a convolutional layer and a sequence of residual convolutional blocks to enhance their representation. Finally, a convolutional layer generates the segmentation map, followed by a sigmoid activation function for binary mask prediction. Experiments are conducted with a batch size of 16 and an input resolution of 448 × 448. The Dice loss is used as the optimization objective, with hyperparameters consistent with the fine-tuning phase. The proposed segmentation head's architecture is shown in Fig. 12, and detailed segmentation performance metrics are provided in Table 10. Our results demonstrate that the Radio DINO models outperform the baseline DINO models by a 2.9% improvement in Dice score, highlighting the efficacy of the learned representations.

7.3. Evaluation across image scales

To assess the robustness of feature representations across different image resolutions, we compare the performance of the Radio DINO base model with the baseline DINO base model. Features are extracted from the CLS token at varying input sizes: 448 × 448, 224 × 224, 112 × 112, and 56 × 56. A logistic regression classifier is trained on these features, with the F1 score serving as the primary evaluation metric. The results show that the Radio DINO base model consistently outperforms the baseline DINO base model across all image scales. Specifically, the Radio DINO base model achieves an average F1 score of 81.12%, compared to 77.82% for the DINO base model. This performance gap highlights the superior generalizability and robustness of the Radio DINO model's learned features, especially when handling varying image resolutions critical in clinical applications.

7.4. Dice score evaluation on Grad-CAM maps

To further evaluate the models' localization capabilities, we compute the Dice score between thresholded Grad-CAM activation maps and the corresponding ground truth segmentation masks. A fixed threshold of 20% is applied to the Grad-CAM outputs to generate binary masks, enabling direct comparison with segmentation annotations. Our results show that the Radio DINO base model outperforms the baseline DINO base model by 2.8%, achieving a Dice score of 13.8%. This improvement indicates the enhanced ability of the Radio DINO model to focus on the most diagnostically relevant regions in the medical images. The performance gain is likely due to the model's refined attention mechanisms, which better localize critical areas while ignoring irrelevant background features. This capability is especially important as the ground truth segmentation masks cover the entire tumoral region (see Fig. 11).

8. Limitations

While Radio DINO demonstrates strong performance across multiple clinical tasks, several limitations must be acknowledged. First, its effectiveness is highly dependent on the quality and diversity of the training data. Despite leveraging the large-scale RadImageNet dataset,

this dataset may not fully encapsulate the variety of medical images encountered in real-world clinical practice. Consequently, the model may struggle with rare or underrepresented cases, limiting its generalization potential in highly specialized medical contexts. Second, although our training configuration consistently outperforms baseline models with an improvement exceeding 1% across all key metrics, this margin, while statistically significant, remains relatively small. However, in medical imaging applications particularly on widely used benchmark datasets such as MedMNISTv2 even minor improvements can be clinically meaningful. Enhanced diagnostic accuracy, even by small margins, can translate into more reliable medical assessments, potentially reducing diagnostic errors and improving patient outcomes. Third, heatmap-based visualizations, such as Grad-CAM, used to explain the model's decisions have inherent limitations. While these tools provide approximate indications of important regions, they may lack granularity and fail to delineate clinically relevant features precisely. Further, heatmaps may emphasize coarse, high-level features that do not necessarily align with the detailed visual cues radiologists rely on. Future studies should explore alternative interpretability methods, such as counterfactual explanations or region-specific attention maps, to better support clinical decision-making. Transparent, trustworthy AI-driven diagnostics necessitate additional validation studies, including clinician-in-the-loop evaluations to assess the practical usability of the visual explanations. Fourth, we only report pure imaging-based experiments throughout our entire set of evaluations. This can be seen as both an application of our model to address the main radiological challenges, but it also represents a significant limitation our model is currently able to interpret and process only imaging data. This approach does not fully emulate the typical workflow of a radiological expert, where the diagnostic process involves a sequence of trials and complementary tests that collectively inform the final diagnosis. Despite this limitation, we plan to expand our experiments in future work by leveraging our Radio DINO model as an image encoder within larger, multimodal environments such as those described in [71,72]. In these settings, our foundation models could serve directly as feature encoders for radiological inputs or, alternatively, enrich their representations through training strategies that integrate multiple modalities like histology or microscopy. However, the latter approach would necessitate different training strategies and, most critically, large and sufficiently diverse datasets. Nevertheless, various methods not exclusively self-supervised learning have demonstrated impressive results toward this goal [73]. Finally, computational demands represent a constraint. Although self-supervised learning reduces the reliance on labeled data, pre-training Radio DINO on large-scale datasets remains computationally intensive, potentially limiting its adoption in resource-constrained healthcare settings. Moreover, the use of Vision Transformers, which serve as the backbone for Radio DINO, entails quadratic computational complexity, thereby increasing resource requirements. To address these limitations, future research should explore both more efficient training paradigms and lightweight model adaptations, such as reduced-complexity transformer architectures or novel self-supervised learning schemes, to facilitate broader accessibility.

9. Discussion

Radio DINO introduces a novel approach to radiomics by leveraging self-supervised learning to enhance feature extraction from medical images. By building upon the DINO and DINOv2 architectures, it captures rich imaging representations without requiring extensive manual feature engineering. This advantage is particularly crucial in medical imaging, where acquiring labeled data is expensive and time-consuming. Our experimental results highlight Radio DINO's ability to outperform existing models across multiple radiology tasks, demonstrating superior generalization across different datasets. This robustness is essential for real-world deployment, where variations in image acquisition protocols and patient demographics can significantly impact model performance. The observed improvements, while sometimes incremental in absolute terms, hold significant clinical implications by enhancing diagnostic reliability. Additionally, the integration of visualization techniques enhances model transparency by providing insights into the learned representations. These explainability features not only increase confidence in the model's decisions but also pave the way for human-AI collaboration in clinical workflows. However, further refinement of these techniques is necessary to ensure that highlighted regions correspond to clinically relevant features recognized by radiologists. Despite these advancements, challenges remain in ensuring that self-supervised models like Radio DINO achieve widespread adoption in clinical practice. Future work should focus on fine-tuning model performance on more diverse datasets, improving interpretability, and addressing computational constraints to enable efficient deployment in real-world settings.

10. Potential clinical applications

Radio DINO holds significant potential for clinical applications, particularly in enhancing diagnostic accuracy, efficiency, and decision support within radiology workflows. Its robust feature extraction capabilities, driven by self-supervised learning, allow it to generalize effectively across diverse medical imaging modalities without the need for extensive labeled datasets. This makes it an ideal candidate for deployment in diagnostic decision support systems, where it can assist radiologists by highlighting regions of interest and suggesting potential pathological findings. For example, in chest radiography, Radio DINO could aid in the early detection of conditions such as pneumonia, lung nodules, or interstitial lung diseases, improving diagnostic confidence and reducing oversight risks. Moreover, its ability to identify subtle imaging features enables effective triage in high-volume clinical environments, prioritizing urgent cases and optimizing workflow efficiency. Beyond diagnostic tasks, Radio DINO's capacity to extract quantitative imaging biomarkers supports applications in disease characterization, treatment response assessment, and prognostic evaluation, particularly in oncology where precise tumor segmentation and heterogeneity analysis are critical. Additionally, its adaptability makes it suitable for deployment in resource-limited settings, democratizing access to advanced radiomics tools where expert radiologists may be scarce. Despite its promise, successful clinical integration of Radio DINO will require addressing challenges related to model interpretability, regulatory compliance, and real-world validation.

11. Conclusions

Radio DINO presents a significant step forward in applying self-supervised learning to radiomics, offering an effective alternative to traditional supervised deep learning approaches. By eliminating the need for extensive labeled datasets while maintaining competitive performance, it enhances the accessibility of AI-driven diagnostic tools in medical imaging. Our findings confirm that Radio DINO not only achieves strong classification results but also exhibits promising generalization capabilities across various imaging modalities and clinical tasks. The improved feature extraction capabilities, combined with visualization-based interpretability, provide a foundation for more transparent and trustworthy AI-assisted diagnostics. Future research should prioritize addressing the identified limitations, particularly in enhancing model interpretability and ensuring adaptability across diverse clinical environments. Collaboration between AI researchers and medical professionals will be essential in refining these models, validating their real-world applicability, and integrating them into routine clinical workflows. By bridging the gap between cutting-edge AI techniques and clinical utility, models like Radio DINO have the potential to transform radiomics, ultimately contributing to improved patient care and medical decision-making.

Acronyms

The following acronyms are used in this manuscript:

DINO	Distillation with No Labels
SSL	Self Supervised Learning
MR	Magnetic Resonance
CT	Computed Tomography
AI	Artificial Intelligence
NLP	Natural Language Processing
ViT	Vision Transformer
EMA	Exponential Moving Average
AUC	Area Under the Curve
TN	True Negatives
FP	False Positives
FN	False Negatives
TP	True Positives
AMP	Automatic Mixed Precision
kNN	k-Nearest Neighbor
PCA	Principal Component Analysis
UMAP	Uniform Manifold Approximation and Projection

CRediT authorship contribution statement

Luca Zedda: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Andrea Loddo:** Writing – review & editing, Writing – original draft, Validation, Supervision, Investigation, Formal analysis, Conceptualization. **Cecilia Di Ruberto:** Writing – review & editing, Writing – original draft, Validation, Supervision, Investigation, Funding acquisition, Formal analysis, Conceptualization.

Ethics statement

This study does not require an Ethics Statement as it does not involve direct experimentation with human subjects or personal data collection. All medical imaging datasets used, including RadImageNet and MedMNISTv2 and BUSI, are publicly available and anonymized, ensuring compliance with ethical guidelines and privacy regulations. No additional patient data was collected or processed beyond these datasets.

Code availability

The code for this study and pretrained models are available at the following GitHub repository: https://github.com/Snarci/Radio-DINO.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Luca Zedda reports financial support was provided by Italian Ministry of University and Research. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We acknowledge financial support under the National Recovery and Resilience Plan (NRRP), Mission 4 Component 2 Investment 1.5 - Call for tender No. 3277 published on December 30, 2021 by the Italian Ministry of University and Research (MUR) funded by the European Union – NextGenerationEU. Project Code ECS0000038 – Project Title eINS Ecosystem of Innovation for Next Generation Sardinia – CUP F53C22000430001- Grant Assignment Decree No. 1056 adopted on June 23, 2022 by the Italian Ministry of University and Research (MUR).

References

[1] B. Koçak, E.Ş. Durmaz, E. Ateş, Ö. Kılıçkesmez, Radiomics with artificial intelligence: a practical guide for beginners, Diagn. Interv. Radiol. 25 (6) (2019) 485.

[2] C. McCague, S. Ramlee, M. Reinius, I. Selby, D. Hulse, P. Piyatissa, V. Bura, M. Crispin-Ortuzar, E. Sala, R. Woitek, Introduction to radiomics for a clinical audience, Clin. Radiol. 78 (2) (2023) 83–98, http://dx.doi.org/10.1016/j.crad.2022.08.149, URL: https://linkinghub.elsevier.com/retrieve/pii/S000992602200705X.

[3] C. Scapicchio, M. Gabelloni, A. Barucci, D. Cioni, L. Saba, E. Neri, A deep look into radiomics, La Radiol. Medica 126 (10) (2021) 1296–1311, http://dx.doi.org/10.1007/s11547-021-01389-x, URL: https://link.springer.com/10.1007/s11547-021-01389-x.

[4] L.R. Wishart, E.C. Ward, G. Galloway, Advances in and applications of imaging and radiomics in head and neck cancer survivorship, Curr. Opin. Otolaryngol. Head Neck Surg. 31 (6) (2023) 368–373, http://dx.doi.org/10.1097/MOO.0000000000000918, URL: https://journals.lww.com/10.1097/MOO.0000000000000918.

[5] H.M. Dragoş, A. Stan, R. Pintican, D. Feier, A. Lebovici, P.-S. Panaitescu, C. Dina, S. Strilciuc, D.F. Muresanu, MRI radiomics and predictive models in assessing ischemic stroke outcome—A systematic review, Diagnostics 13 (5) (2023) 857.

[6] W. Zhang, Y. Guo, Q. Jin, Radiomics and its feature selection: A review, Symmetry 15 (10) (2023) 1834, http://dx.doi.org/10.3390/sym15101834, URL: https://www.mdpi.com/2073-8994/15/10/1834.

[7] M. Caron, H. Touvron, I. Misra, H. Jegou, J. Mairal, P. Bojanowski, A. Joulin, Emerging properties in self-supervised vision transformers, in: 2021 IEEE/CVF International Conference on Computer Vision, ICCV, IEEE, Montreal, QC, Canada, 2021, pp. 9630–9640, http://dx.doi.org/10.1109/ICCV48922.2021.00951, URL: https://ieeexplore.ieee.org/document/9709990/.

[8] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, et al., Dinov2: Learning robust visual features without supervision, 2023, arXiv preprint arXiv:2304.07193.

[9] T. Darcet, M. Oquab, J. Mairal, P. Bojanowski, Vision transformers need registers, 2024, http://dx.doi.org/10.48550/arXiv.2309.16588, URL: http://arxiv.org/abs/2309.16588, arXiv:2309.16588 [cs].

[10] M. Assran, Q. Duval, I. Misra, P. Bojanowski, P. Vincent, M. Rabbat, Y. LeCun, N. Ballas, Self-supervised learning from images with a joint-embedding predictive architecture, in: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, IEEE, Vancouver, BC, Canada, 2023, pp. 15619–15629, http://dx.doi.org/10.1109/CVPR52729.2023.01499, URL: https://ieeexplore.ieee.org/document/10205476/.

[11] Y. Fang, W. Wang, B. Xie, Q. Sun, L. Wu, X. Wang, T. Huang, X. Wang, Y. Cao, EVA: Exploring the limits of masked visual representation learning at scale, in: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, IEEE, Vancouver, BC, Canada, 2023, pp. 19358–19369, http://dx.doi.org/10.1109/CVPR52729.2023.01855, URL: https://ieeexplore.ieee.org/document/10203681/.

[12] K. He, X. Chen, S. Xie, Y. Li, P. Dollar, R. Girshick, Masked autoencoders are scalable vision learners, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, IEEE, New Orleans, LA, USA, 2022, pp. 15979–15988, http://dx.doi.org/10.1109/CVPR52688.2022.01553, URL: https://ieeexplore.ieee.org/document/9879206/.

[13] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language supervision, in: Proceedings of the 38th International Conference on Machine Learning, PMLR, (ISSN: 2640-3498) 2021, pp. 8748–8763, URL: https://proceedings.mlr.press/v139/radford21a.html.

[14] X. Liu, J. Liang, J. Zhang, Z. Qian, P. Xing, T. Chen, S. Yang, C. Chukwudi, L. Qiu, D. Liu, J. Zhao, Advancing hierarchical neural networks with scale-aware pyramidal feature learning for medical image dense prediction, Comput. Methods Programs Biomed. 265 (2025) 108705, http://dx.doi.org/10.1016/j.cmpb.2025.108705, URL: https://www.sciencedirect.com/science/article/pii/S0169260725001221.

[15] Z. Li, H. Li, A.L. Ralescu, J.R. Dillman, N.A. Parikh, L. He, A novel collaborative self-supervised learning method for radiomic data, NeuroImage 277 (2023) 120229, http://dx.doi.org/10.1016/j.neuroimage.2023.120229, URL: https://www.sciencedirect.com/science/article/pii/S1053811923003804.

[16] C. Ouchicha, O. Ammor, M. Meknassi, CVDNet: A novel deep learning architecture for detection of coronavirus (Covid-19) from chest x-ray images, Chaos Solitons Fractals 140 (2020) 110245, http://dx.doi.org/10.1016/j.chaos.2020.110245, URL: https://www.sciencedirect.com/science/article/pii/S096007792030641X.

[17] S. Huang, Y. Ge, D. Liu, M. Hong, J. Zhao, A.C. Loui, Rethinking copy-paste for consistency learning in medical image segmentation, IEEE Trans. Image Process. 34 (2025) 1060–1074, http://dx.doi.org/10.1109/TIP.2025.3536208, URL: https://ieeexplore.ieee.org/document/10871927.

[18] G.I. Okolo, S. Katsigiannis, N. Ramzan, IEViT: An enhanced vision transformer architecture for chest X-ray image classification, Comput. Methods Programs Biomed. 226 (2022) 107141, http://dx.doi.org/10.1016/j.cmpb.2022.107141, URL: https://www.sciencedirect.com/science/article/pii/S0169260722005223.

[19] S. Pai, D. Bontempi, I. Hadzic, V. Prudente, M. Sokač, T.L. Chaunzwa, S. Bernatz, A. Hosny, R.H. Mak, N.J. Birkbak, H.J.W.L. Aerts, Foundation model for cancer imaging biomarkers, Nat. Mach. Intell. 6 (3) (2024) 354–367, http://dx.doi.org/10.1038/s42256-024-00807-9, URL: https://www.nature.com/articles/s42256-024-00807-9, Publisher: Nature Publishing Group.

[20] M. Waseem Sabir, M. Farhan, N.S. Almalki, M.M. Alnfiai, G.A. Sampedro, Frontiers | FibroVit—Vision transformer-based framework for detection and classification of pulmonary fibrosis from chest CT images. http://dx.doi.org/10.3389/fmed.2023.1282200, URL: https://www.frontiersin.org/journals/medicine/articles/10.3389/fmed.2023.1282200/full.

[21] J. Ding, S. Ma, L. Dong, X. Zhang, S. Huang, W. Wang, N. Zheng, F. Wei, LongNet: Scaling transformers to 1,000,000,000 tokens, 2023, http://dx.doi.org/10.48550/arXiv.2307.02486, URL: http://arxiv.org/abs/2307.02486, arXiv:2307.02486 [cs].

[22] C. Liu, H. Kiryu, 3D medical axial transformer: A lightweight transformer model for 3D brain tumor segmentation, 2024, URL: https://proceedings.mlr.press/v227/liu24b.html.

[23] S. Perera, S. Adhikari, A. Yilmaz, Pocformer: A lightweight transformer architecture for detection of Covid-19 using point of care ultrasound, in: 2021 IEEE International Conference on Image Processing, ICIP, (ISSN: 2381-8549) 2021, pp. 195–199, http://dx.doi.org/10.1109/ICIP42928.2021.9506353, URL: https://ieeexplore.ieee.org/abstract/document/9506353.

[24] Y. Zhang, H. Dong, N. Konz, H. Gu, M.A. Mazurowski, Lightweight transformer backbone for medical object detection, 2022, http://dx.doi.org/10.1007/978-3-031-17979-2_5, URL: https://link.springer.com/chapter/10.1007/978-3-031-17979-2_5.

[25] Z. Yang, H. Zhu, R. Zhang, H. Zhang, J. Wang, C. Wang, M. Chen, F.-F. Yin, Embedding radiomics into vision transformers for multimodal medical image classification, 2025, http://dx.doi.org/10.48550/arXiv.2504.10916, URL: http://arxiv.org/abs/2504.10916, arXiv:2504.10916 [physics].

[26] W. Chen, M. Ayoub, M. Liao, R. Shi, M. Zhang, F. Su, Z. Huang, Y. Li, Y. Wang, K.K.L. Wong, A fusion of VGG-16 and ViT models for improving bone tumor classification in computed tomography, J. Bone Oncol. 43 (2023) 100508, http://dx.doi.org/10.1016/j.jbo.2023.100508, URL: https://www.sciencedirect.com/science/article/pii/S2212137423000416.

[27] J. Qiu, J. Mitra, S. Ghose, C. Dumas, J. Yang, B. Sarachan, M.A. Judson, A multichannel CT and radiomics-guided CNN-ViT (RadCT-CNNViT) ensemble network for diagnosis of pulmonary sarcoidosis, Diagnostics 14 (10) (2024) 1049, http://dx.doi.org/10.3390/diagnostics14101049, URL: https://www.mdpi.com/2075-4418/14/10/1049, Number: 10 Publisher: Multidisciplinary Digital Publishing Institute.

[28] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186, http://dx.doi.org/10.18653/v1/N19-1423, URL: https://aclanthology.org/N19-1423.

[29] J. Zhou, Y. Chen, Z. Hong, W. Chen, Y. Yu, T. Zhang, H. Wang, C. Zhang, Z. Zheng, Training and serving system of foundation models: A comprehensive survey, IEEE Open J. Comput. Soc. 5 (2024) 107–119, http://dx.doi.org/10.1109/OJCS.2024.3380828, URL: https://ieeexplore.ieee.org/document/10478189/.

[30] C. Zhou, Q. Li, C. Li, J. Yu, Y. Liu, G. Wang, K. Zhang, C. Ji, Q. Yan, L. He, H. Peng, J. Li, J. Wu, Z. Liu, P. Xie, C. Xiong, J. Pei, P.S. Yu, L. Sun, A comprehensive survey on pretrained foundation models: A history from BERT to ChatGPT, 2023, URL: http://arxiv.org/abs/2302.09419, arXiv:2302.09419 [cs].

[31] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, G. Hinton, Big self-supervised models are strong semi-supervised learners, 2020, http://dx.doi.org/10.48550/arXiv.2006.10029, URL: http://arxiv.org/abs/2006.10029, arXiv:2006.10029 [cs, stat].

[32] X. Chen, H. Fan, R. Girshick, K. He, Improved baselines with momentum contrastive learning, 2020, http://dx.doi.org/10.48550/arXiv.2003.04297, URL: http://arxiv.org/abs/2003.04297, arXiv:2003.04297 [cs].

[33] C. Zhang, H. Zheng, Y. Gu, Dive into the details of self-supervised learning for medical image analysis, Med. Image Anal. 89 (2023) 102879, http://dx.doi.org/10.1016/j.media.2023.102879, URL: https://www.sciencedirect.com/science/article/pii/S1361841523001391.

[34] M. Moor, O. Banerjee, Z.S.H. Abad, H.M. Krumholz, J. Leskovec, E.J. Topol, P. Rajpurkar, Foundation models for generalist medical artificial intelligence, Nat. 616 (7956) (2023) 259–265, http://dx.doi.org/10.1038/s41586-023-05881-4, URL: https://www.nature.com/articles/s41586-023-05881-4, Publisher: Nature Publishing Group.

[35] B. Azad, R. Azad, S. Eskandari, A. Bozorgpour, A. Kazerouni, I. Rekik, D. Merhof, Foundational models in medical imaging: A comprehensive survey and future vision, 2023, URL: http://arxiv.org/abs/2310.18689, arXiv:2310.18689 [cs].

[36] S. Zhang, D. Metaxas, On the challenges and perspectives of foundation models for medical image analysis, Med. Image Anal. 91 (2024) 102996, http://dx.doi.org/10.1016/j.media.2023.102996, URL: https://linkinghub.elsevier.com/retrieve/pii/S1361841523002566.

[37] J.P. Huix, A.R. Ganeshan, J.F. Haslum, M. Söderberg, C. Matsoukas, K. Smith, Are natural domain foundation models useful for medical image classification? in: 2024 IEEE/CVF Winter Conference on Applications of Computer Vision, WACV, IEEE, Waikoloa, HI, USA, 2024, pp. 7619–7628, http://dx.doi.org/10.1109/WACV57701.2024.00746, URL: https://ieeexplore.ieee.org/document/10483777/.

[38] A. Yan, J. McAuley, X. Lu, J. Du, E.Y. Chang, A. Gentili, C.-N. Hsu, RadBERT: Adapting transformer-based language models to radiology, Radiol.: Artif. Intell. 4 (4) (2022) e210258.

[39] H.-Y. Zhou, S. Yu, C. Bian, Y. Hu, K. Ma, Y. Zheng, Comparing to learn: Surpassing ImageNet pretraining on radiographs by comparing image representations, in: Medical Image Computing and Computer Assisted Intervention – MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I, Springer-Verlag, Berlin, Heidelberg, 2020, pp. 398–407, http://dx.doi.org/10.1007/978-3-030-59710-8_39.

[40] H. Sowrirajan, J. Yang, A.Y. Ng, P. Rajpurkar, MoCo-CXR: MoCo pretraining improves representation and transferability of chest X-ray models, 2021, URL: http://arxiv.org/abs/2010.05352, arXiv:2010.05352 [cs].

[41] S.-C. Huang, A. Pareek, M. Jensen, M.P. Lungren, S. Yeung, A.S. Chaudhari, Self-supervised learning for medical image classification: a systematic review and implementation guidelines, Npj Digit. Med. 6 (1) (2023) 1–16, http://dx.doi.org/10.1038/s41746-023-00811-0, URL: https://www.nature.com/articles/s41746-023-00811-0, Publisher: Nature Publishing Group.

[42] W.-C. Wang, E. Ahn, D. Feng, J. Kim, A review of predictive and contrastive self-supervised learning for medical images, Mach. Intell. Res. 20 (4) (2023) 483–513, http://dx.doi.org/10.1007/s11633-022-1406-4.

[43] V. Koch, S.J. Wagner, S. Kazeminia, E. Sancar, M. Hehr, J. Schnabel, T. Peng, C. Marr, DinoBloom: A foundation model for generalizable cell embeddings in hematology, 2024, http://dx.doi.org/10.48550/arXiv.2404.05022, URL: http://arxiv.org/abs/2404.05022, arXiv:2404.05022 [cs].

[44] T. Ding, S.J. Wagner, A.H. Song, R.J. Chen, M.Y. Lu, A. Zhang, A.J. Vaidya, G. Jaume, M. Shaban, A. Kim, D.F.K. Williamson, B. Chen, C. Almagro-Perez, P. Doucet, S. Sahai, C. Chen, D. Komura, A. Kawabe, S. Ishikawa, G. Gerber, T. Peng, L.P. Le, F. Mahmood, Multimodal whole slide foundation model for pathology, 2024, http://dx.doi.org/10.48550/arXiv.2411.19666, URL: http://arxiv.org/abs/2411.19666, arXiv:2411.19666 [eess].

[45] X. Wang, J. Zhao, E. Marostica, W. Yuan, J. Jin, J. Zhang, R. Li, H. Tang, K. Wang, Y. Li, F. Wang, Y. Peng, J. Zhu, J. Zhang, C.R. Jackson, J. Zhang, D. Dillon, N.U. Lin, L. Sholl, T. Denize, D. Meredith, K.L. Ligon, S. Signoretti, S. Ogino, J.A. Golden, M.P. Nasrallah, X. Han, S. Yang, K.-H. Yu, A pathology foundation model for cancer diagnosis and prognosis prediction, Nat. 634 (8035) (2024) 970–978, http://dx.doi.org/10.1038/s41586-024-07894-z, URL: https://www.nature.com/articles/s41586-024-07894-z, Publisher: Nature Publishing Group.

[46] H. Xu, N. Usuyama, J. Bagga, S. Zhang, R. Rao, T. Naumann, C. Wong, Z. Gero, J. González, Y. Gu, Y. Xu, M. Wei, W. Wang, S. Ma, F. Wei, J. Yang, C. Li, J. Gao, J. Rosemon, T. Bower, S. Lee, R. Weerasinghe, B.J. Wright, A. Robicsek, B. Piening, C. Bifulco, S. Wang, H. Poon, A whole-slide foundation model for digital pathology from real-world data, Nat. 630 (8015) (2024) 181–188, http://dx.doi.org/10.1038/s41586-024-07441-w, URL: https://www.nature.com/articles/s41586-024-07441-w, Publisher: Nature Publishing Group.

[47] R.J. Chen, T. Ding, M.Y. Lu, D.F.K. Williamson, G. Jaume, A.H. Song, B. Chen, A. Zhang, D. Shao, M. Shaban, M. Williams, L. Oldenburg, L.L. Weishaupt, J.J. Wang, A. Vaidya, L.P. Le, G. Gerber, S. Sahai, W. Williams, F. Mahmood, Towards a general-purpose foundation model for computational pathology, Nature Med. 30 (3) (2024) 850–862, http://dx.doi.org/10.1038/s41591-024-02857-3, URL: https://www.nature.com/articles/s41591-024-02857-3, Publisher: Nature Publishing Group.

[48] J. Qiu, W. Yuan, K. Lam, The application of multimodal large language models in medicine, Lancet Reg. Heal. – West. Pac. 45 (2024) http://dx.doi.org/10.1016/j.lanwpc.2024.101048.

[49] K. Lam, J. Qiu, Foundation models: the future of surgical artificial intelligence? Br. J. Surg. 111 (4) (2024) znae090, http://dx.doi.org/10.1093/bjs/znae090.

[50] C. Liu, Y. Tian, W. Chen, Y. Song, Y. Zhang, Bootstrapping large language models for radiology report generation, Proc. AAAI Conf. Artif. Intell. 38 (17) (2024) 18635–18643, http://dx.doi.org/10.1609/aaai.v38i17.29826, URL: https://ojs.aaai.org/index.php/AAAI/article/view/29826, Number: 17.

[51] S.-C. Huang, L. Shen, M.P. Lungren, S. Yeung, GLoRIA: A multimodal global-local representation learning framework for label-efficient medical image recognition, in: 2021 IEEE/CVF International Conference on Computer Vision, ICCV, (ISSN: 2380-7504) 2021, pp. 3922–3931, http://dx.doi.org/10.1109/ICCV48922.2021.00391, URL: https://ieeexplore.ieee.org/document/9710099.

[52] C. Li, C. Wong, S. Zhang, N. Usuyama, H. Liu, J. Yang, T. Naumann, H. Poon, J. Gao, LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day, 2023, URL: http://arxiv.org/abs/2306.00890, arXiv:2306.00890 [cs].

[53] Y. Xie, J. Zhang, L. Liu, H. Wang, Y. Ye, J. Verjans, Y. Xia, ReFs: A hybrid pre-training paradigm for 3D medical image segmentation, Med. Image Anal. 91 (2024) 103023, http://dx.doi.org/10.1016/j.media.2023.103023, URL: https://www.sciencedirect.com/science/article/pii/S1361841523002839.

[54] Y. Ye, J. Zhang, Z. Chen, Y. Xia, CADS: A self-supervised learner via cross-modal alignment and deep self-distillation for CT volume segmentation, IEEE Trans. Med. Imaging 44 (1) (2025) 118–129, http://dx.doi.org/10.1109/TMI.2024.3431916, URL: https://ieeexplore.ieee.org/document/10605840, Conference Name: IEEE Transactions on Medical Imaging.

[55] X. Mei, Z. Liu, P.M. Robson, B. Marinelli, M. Huang, A. Doshi, A. Jacobi, C. Cao, K.E. Link, T. Yang, et al., RadImageNet: an open radiologic deep learning research dataset for effective transfer learning, Radiol.: Artif. Intell. 4 (5) (2022) e210315.

[56] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, ImageNet: A large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, (ISSN: 1063-6919) 2009, pp. 248–255, http://dx.doi.org/10.1109/CVPR.2009.5206848, URL: https://ieeexplore.ieee.org/document/5206848/?arnumber=5206848.

[57] J. Yang, R. Shi, D. Wei, Z. Liu, L. Zhao, B. Ke, H. Pfister, B. Ni, MedMNIST v2-a large-scale lightweight benchmark for 2D and 3D biomedical image classification, Sci. Data 10 (1) (2023) 41.

[58] W. Al-Dhabyani, M. Gomaa, H. Khaled, A. Fahmy, Dataset of breast ultrasound images, Data Brief 28 (2020) 104863, http://dx.doi.org/10.1016/j.dib.2019.104863, URL: https://www.sciencedirect.com/science/article/pii/S2352340919312181.

[59] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L.u. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, Vol. 30, Curran Associates, Inc., 2017, p. 11, URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[60] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, in: ICLR, 2021.

[61] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A.C. Berg, W.-Y. Lo, P. Dollár, R. Girshick, Segment anything, in: 2023 IEEE/CVF International Conference on Computer Vision, ICCV, IEEE, Paris, France, 2023, pp. 3992–4003, http://dx.doi.org/10.1109/ICCV51070.2023.00371, URL: https://ieeexplore.ieee.org/document/10378323/.

[62] X. Zhai, B. Mustafa, A. Kolesnikov, L. Beyer, Sigmoid loss for language image pre-training, in: 2023 IEEE/CVF International Conference on Computer Vision, ICCV, IEEE, Paris, France, 2023, pp. 11941–11952, http://dx.doi.org/10.1109/ICCV51070.2023.01100, URL: https://ieeexplore.ieee.org/document/10377550/.

[63] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, IEEE, Las Vegas, NV, USA, 2016, pp. 770–778, http://dx.doi.org/10.1109/CVPR.2016.90, URL: http://ieeexplore.ieee.org/document/7780459/.

[64] T. Darcet, M. Oquab, J. Mairal, P. Bojanowski, Vision transformers need registers, in: The Twelfth International Conference on Learning Representations, 2024, p. 21, URL: https://openreview.net/forum?id=2dnO3LLiJ1.

[65] O.N. Manzari, H. Ahmadabadi, H. Kashiani, S.B. Shokouhi, A. Ayatollahi, MedViT: A robust vision transformer for generalized medical image classification, Comput. Biol. Med. 157 (2023) 106791, http://dx.doi.org/10.1016/j.compbiomed.2023.106791, URL: http://arxiv.org/abs/2302.09462, arXiv:2302.09462 [cs].

[66] R. Zhong, D. Ghosh, D. Klein, J. Steinhardt, Are larger pretrained language models uniformly better? comparing performance at the instance level, in: C. Zong, F. Xia, W. Li, R. Navigli (Eds.), Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Association for Computational Linguistics, Online, 2021, pp. 3813–3827, http://dx.doi.org/10.18653/v1/2021.findings-acl.334, URL: https://aclanthology.org/2021.findings-acl.334.

[67] A. Singh, P. Ven, C. Eising, P. Denny, Dynamic filter application in graph convolutional networks for enhanced spectral feature analysis and class discrimination in medical imaging, IEEE Access PP (2024) http://dx.doi.org/10.1109/ACCESS.2024.3444042, 1–1.

[68] R. Schäfer, T. Nicke, H. Höfener, A. Lange, D. Merhof, F. Feuerhake, V. Schulz, J. Lotz, F. Kiessling, Overcoming data scarcity in biomedical imaging with a foundational multi-task model, Nat. Comput. Sci. 4 (2024) 1–15, http://dx.doi.org/10.1038/s43588-024-00662-z.

[69] Z. Lai, J. Wu, S. Chen, Y. Zhou, N. Hovakimyan, Residual-based language models are free boosters for biomedical imaging tasks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2024, pp. 5086–5096.

[70] Z. Chen, Y. Duan, W. Wang, J. He, T. Lu, J. Dai, Y. Qiao, Vision transformer adapter for dense predictions, 2022, arXiv preprint arXiv:2205.08534.

[71] Z. Li, Y. Jiang, M. Lu, R. Li, Y. Xia, Survival prediction via hierarchical multimodal co-attention transformer: A computational histology-radiology solution, IEEE Trans. Med. Imaging 42 (9) (2023) 2678–2689, http://dx.doi.org/10.1109/TMI.2023.3263010, URL: https://ieeexplore.ieee.org/document/10086562/.

[72] R.S. Vanguri, J. Luo, A.T. Aukerman, J.V. Egger, C.J. Fong, N. Horvat, A. Pagano, J.d.B. Araujo-Filho, L. Geneslaw, H. Rizvi, R. Sosa, K.M. Boehm, S.-R. Yang, F.M. Bodd, K. Ventura, T.J. Hollmann, M.S. Ginsberg, J. Gao, R. Vanguri, M.D. Hellmann, J.L. Sauter, S.P. Shah, Multimodal integration of radiology, pathology and genomics for prediction of response to PD-(L)1 blockade in patients with non-small cell lung cancer, Nat. Cancer 3 (10) (2022) 1151–1164, http://dx.doi.org/10.1038/s43018-022-00416-8, URL: https://www.nature.com/articles/s43018-022-00416-8, Publisher: Nature Publishing Group.

[73] Z. Ji, Y. Ge, C. Chukwudi, K. U, S.M. Zhang, Y. Peng, J. Zhu, H. Zaki, X. Zhang, S. Yang, X. Wang, Y. Chen, J. Zhao, Counterfactual bidirectional co-attention transformer for integrative histology-genomic cancer risk stratification, IEEE J. Biomed. Heal. Inform. (2025) 1–13, http://dx.doi.org/10.1109/JBHI.2025.3548048, URL: https://ieeexplore.ieee.org/document/10910138/.