scientific reports

Check for updates

OPEN

# Medical slice transformer for improved diagnosis and explainability on 3D medical images with DINOv2

Gustav Müller-Franzes[1]✉, Firas Khader[1], Robert Siepmann[1], Tianyu Han[1], Jakob Nikolas Kather[2,3,4], Sven Nebelung[1,5] & Daniel Truhn[1,5]

**Magnetic Resonance Imaging (MRI) and Computed Tomography (CT) are essential clinical cross-sectional imaging techniques for diagnosing complex conditions. However, large 3D datasets with annotations for deep learning are scarce. While methods like DINOv2 are encouraging for 2D image analysis, these methods have not been applied to 3D medical images. Furthermore, deep learning models often lack explainability due to their "black-box" nature. This study aims to extend 2D self-supervised models, specifically DINOv2, to 3D medical imaging while evaluating their potential for explainable outcomes. We introduce the Medical Slice Transformer (MST) framework to adapt 2D self-supervised models for 3D medical image analysis. MST combines a Transformer architecture with a 2D feature extractor, i.e., DINOv2. We evaluate its diagnostic performance against a 3D convolutional neural network (3D ResNet) across three clinical datasets: breast MRI (651 patients), chest CT (722 patients), and knee MRI (1199 patients). Both methods were tested for diagnosing breast cancer, predicting lung nodule dignity, and detecting meniscus tears. Diagnostic performance was assessed by calculating the Area Under the Receiver Operating Characteristic Curve (AUC). Explainability was evaluated through a radiologist's qualitative comparison of saliency maps based on slice and lesion correctness. P-values were calculated using Delong's test. MST achieved higher AUC values compared to ResNet across all three datasets: breast (0.94 ± 0.01 vs. 0.91 ± 0.02, $P = 0.02$), chest (0.95 ± 0.01 vs. 0.92 ± 0.02, $P = 0.13$), and knee (0.85 ± 0.04 vs. 0.69 ± 0.05, $P = 0.001$). Saliency maps were consistently more precise and anatomically correct for MST than for ResNet. Self-supervised 2D models like DINOv2 can be effectively adapted for 3D medical imaging using MST, offering enhanced diagnostic accuracy and explainability compared to convolutional neural networks.**

Deep learning (DL) has demonstrated significant potential in medical imaging for diagnosis and analysis[1,2]. However, its integration into clinical practice is hindered by the need for large, annotated datasets and limited model interpretability. Generating medical annotations is time-consuming and requires expert input[3,4]. Furthermore, the "black-box" nature of DL models raises concerns about trustworthiness in clinical settings[5–7].

To address the challenge of data scarcity, self-supervised learning has emerged as a powerful paradigm, enabling models to learn from vast quantities of unlabeled data, often in conjunction with smaller labeled datasets to enhance performance[16]. Foundation models like DINOv2[17] use self-supervised learning to extract features from vast amounts of unlabeled data. Despite being trained on natural images, these models have shown promise in transferring features to medical imaging tasks[18–24]. For instance, Huix et al.[8] found that DINOv2 outperformed other foundation models and even in-domain self-supervised ResNet152 in tasks like fundoscopy, mammography, dermoscopy, and chest radiography. Similarly, Truong et al.[9] and Nielsen et al.[10] demonstrated DINOv2's effectiveness on microscopic and endoscopic images. Huang et al.[11] reported that DINOv2 generally outperformed VGG and ResNet50 models pre-trained on ImageNet for chest radiographic, funduscopic, and dermoscopic images but not for brain Magnetic Resonance Imaging (MRI) slices.

[1]Department of Diagnostic and Interventional Radiology, University Hospital Aachen, Aachen, Germany. [2]Else Kroener Fresenius Center for Digital Health, Technical University Dresden, Dresden, Germany. [3]Department of Medicine I, University Hospital Dresden, Dresden, Germany. [4]National Center for Tumor Diseases (NCT), University Hospital Heidelberg, Heidelberg, Germany. [5]Sven Nebelung and Daniel Truhn contributed equally to this work. ✉email: gumueller@ukaachen.de

To enhance model explainability, techniques like saliency maps and Gradient-weighted Class Activation Mapping (Grad-CAM)[12] have been used to visualize network attention and provide insights into model decisions. However, these methods often lack accuracy, especially in 3D medical imaging[13,14]. Transformers offer an alternative through their inherent attention mechanisms[15,16]. Unlike convolutional models requiring post-hoc visualization, Transformers provide inherent attention matrices highlighting the most relevant input features. Studies have shown that attention-based saliency maps improve interpretability in 2D medical imaging[17–19].

Despite these advancements in leveraging self-supervised learning for 2D medical tasks and the potential of Transformers for explainability, a research gap persists: the application of 2D self-supervised foundation models like DINOv2 to 3D medical imaging, and the exploitation of their inherent attention mechanisms for enhanced explainability.

The present study aims to bridge this gap by introducing the Medical Slice Transformer (MST), a novel approach that adapts 2D self-supervised models like DINOv2 for 3D medical image analysis. The MST framework operates by first employing a pre-trained 2D self-supervised model (DINOv2) to extract rich, high-quality features from individual 2D slices of a 3D medical volume. These slice-wise features are then processed by a Transformer architecture. The Transformer component is used to aggregate these features, modeling inter-slice dependencies and learning the contextual relationships across the entire 3D volume to arrive at a diagnostic decision (Fig. 1). This approach contrasts with conventional methods that might involve training 3D models from scratch with limited data or simpler forms of 2D feature aggregation that do not leverage the rich, representative features from self-supervised learning models like DINOv2, nor incorporate the attention mechanisms of Transformer architectures for model explainability.

We hypothesized that the MST framework would outperform standard 3D convolutional neural networks (such as ResNet) in terms of diagnostic accuracy across a set of clinical cross-sectional Computed Tomography (CT) and MRI datasets. Furthermore, we hypothesized that by leveraging the Transformer's attention mechanism applied to the DINOv2-derived slice features, MST would provide better localization information for identifying specific image findings, such as breast lesions, lung nodules, and meniscus tears, compared to ResNet.

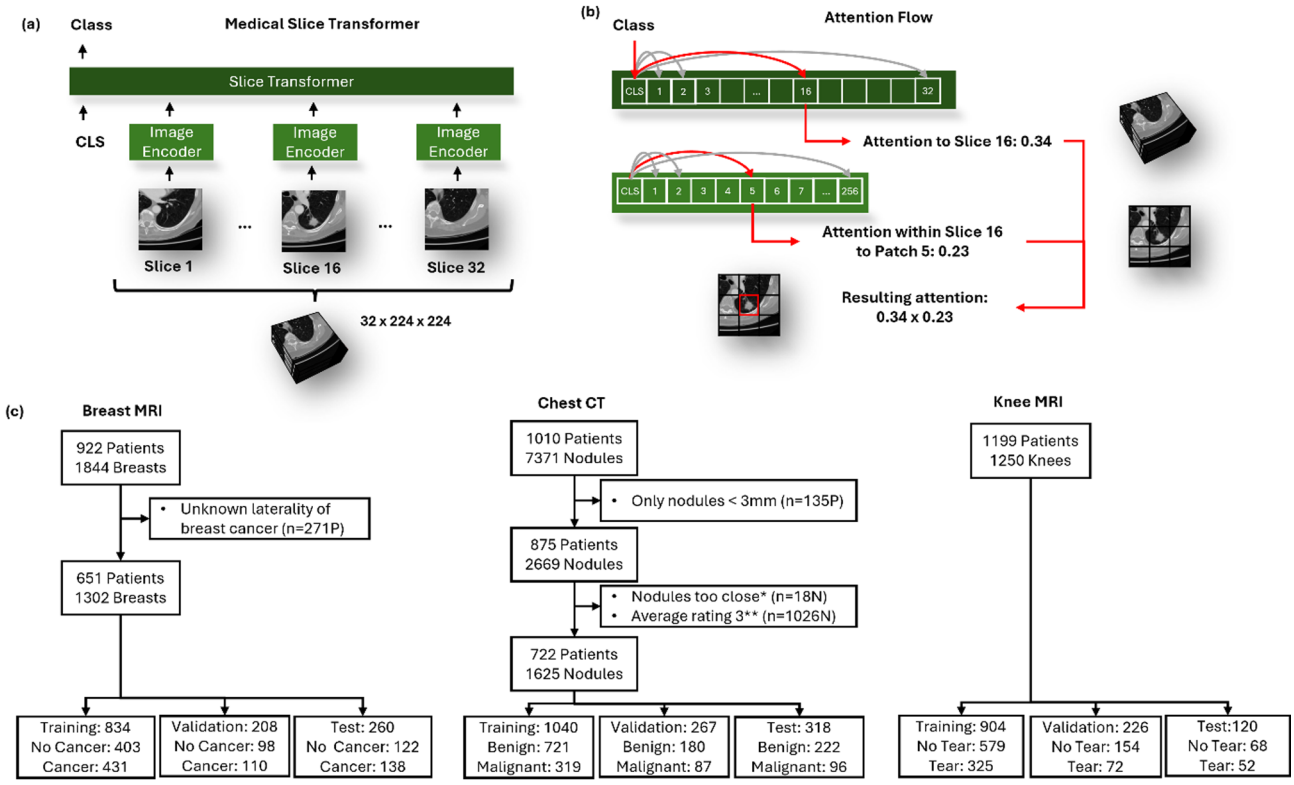



**Fig. 1**. Overview of the model architecture, attention flow, and study workflow. (**a**) The Medical Slice Transformer framework processes individual MRI or CT slices using 2D image encoders, such as DINOv2, and then passes the encoded outputs through the Slice Transformer for downstream classification tasks. (**b**) Visualization of attention mechanisms showing how the Slice Transformer assigns attention to specific slices and how within-slice attention is further refined to specific patches, resulting in a combined attention map highlighting regions of interest in the input volume. (**c**) Study flow diagram of the Breast MRI dataset (Duke-Breast-Cancer-MRI), Chest CT dataset (Lung Image Database Consortium and Image Database Resource Initiative), and Knee MRI dataset (MRNet). Abbreviations: *Proximity of the nodules made them unresolvable by the pylidc library. **Radiologists rated the likelihood of malignancy on a 5-point scale ranging from "1: highly unlikely" to "5: highly suspicious"; thus, an average rating of "3" indicated unclear dignity. *CLS* classification token, *P* Patient, *N* Nodules.

The primary contributions of this study are as follows:

- Development of MST: We introduce MST, a technique designed to use self-supervised models like DINOv2 for 3D medical image analysis.
- Enhanced explainability: By leveraging the Transformer's attention mechanism, we demonstrate how model explainability can be improved, providing more insight into the decision-making process of DL models in 3D medical imaging contexts.

## Results

### Patient characteristics

Three publicly available datasets were included: (i) breast MRI "Duke-Breast-Cancer-MRI" (DUKE)[20], (ii) chest CT "Lung Image Database Consortium and Image Database Resource Initiative" (LIDC-IDRI)[21], and (iii) knee MRI "MRNet"[22]. Inclusion criteria were met by 651, 722, and 1199 patients for the breast MRI, chest CT, and knee MRI datasets, respectively (Fig. 1c). The DUKE dataset comprised contrast-enhanced breast MR subtraction images from cancer patients and was used to evaluate the models' performance in breast cancer detection. The LIDC-IDRI dataset included chest CT scans with lung nodule annotations by four radiologists, allowing assessment of the models' ability to detect and analyze the dignity of lung nodules. The MRNet dataset contained knee MRI studies with meniscal tears, enabling evaluation of the models' effectiveness in diagnosing meniscal tears.

### MST outperforms standard CNN architectures

We compared the performance of a 3D ResNet50[23] with the proposed MST architecture using DINOv2 as the image encoder by calculating the Area Under the Receiver Operating Characteristic Curve (AUC) for each dataset (Fig. 2). MST exhibited higher AUC values compared to the 3D ResNet on all datasets: breast MRI: $0.94 \pm 0.01$ vs. $0.91 \pm 0.02$ ($P = 0.02$), chest CT: $0.95 \pm 0.01$ vs. $0.92 \pm 0.02$ ($P = 0.13$), and knee MRI: $0.85 \pm 0.04$ vs. $0.69 \pm 0.05$ ($P = 0.001$).

When systematically varying the reference model's complexity, we found that ResNet18 or ResNet101 did not significantly improve performance over ResNet50 (Supplementary Table S1). ResNet18, ResNet50, and ResNet101 have 33 million, 46 million, and 117 million parameters, respectively. In contrast, the MST model has 23 million parameters, with 1 million parameters belonging to the Transformer part and 22 million to the DINOv2 part.

### MST provides better model explainability

A major challenge with deep learning models in medical imaging is their lack of transparency, often called the "black-box" problem. For reference, we included the traditional Gradient-weighted Class Activation Mapping (Grad-CAM) to visualize the attention of the ResNet model.

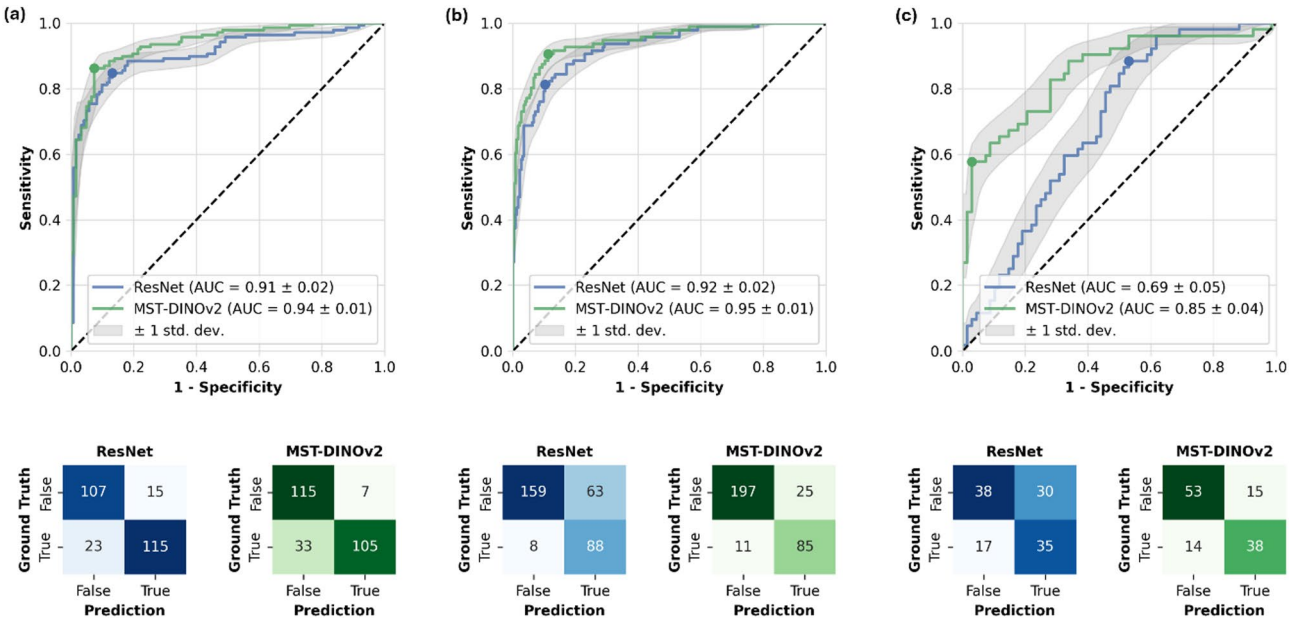



**Fig. 2**. Classification performance as a function of model architecture and dataset. Shown are the Area Under the Receiver Operating Characteristic Curve (AUC) values and confusion matrices for the Transformer-based self-supervised model (MST-DINOv2) and the standard convolutional neural network (ResNet) in the diagnosis of breast cancer in MRI (DUKE dataset, (**a**) dignity assessment of lung nodules in CT (LIDC-IDRI dataset, (**b**), and presence of meniscal tears (MRNet dataset, **c**).

To complement the quantitative analysis, we evaluated the saliency maps provided by both methods. A radiologist (R.S.) with five years of clinical experience performed a blinded evaluation of 50 randomly selected saliency maps from each dataset, rating the models based on two key criteria (Fig. 3):

(i) Slice correctness—does the saliency map highlight the slice(s) containing the lesion?
(ii) Lesion correctness—does the saliency map accurately point to the core of the lesion?

Irrespective of the clinical dataset, the saliency maps generated by the 3D ResNet were often distributed across multiple slices that did not contain lesions. Compared to the ground truth (Figs. 4a, 5a and 6a), the saliency maps provided minimal information about the precise location of the different lesions (Figs. 4b, 5b and 6b). The radiologist rated 37 of the provided 150 saliency maps as correctly highlighting the slice containing the lesion, and none (in any dataset) as accurately pointing to the core of the lesion (Table 1).

In contrast, the slice-wise attention maps produced by the MST framework effectively highlighted relevant slices (Figs. 4c, 5c and 6c) and, when combined with the within-slice attention, pointed out the core of the lesions more reliably (Figs. 4d, 5d and 6d). The radiologist rated 136 of the provided 150 saliency maps as correctly highlighting the slice(s) containing the lesion, and 57 of 150 as accurately pointing to the core of the lesion.

## Ablation experiments

We conducted ablation studies to analyze the optimal architecture for the Medical Slice Transformer (MST) framework. First, we evaluated the importance of the Transformer for aggregating slices on the diagnostic performance. Second, we explored different architectures and associated parameters for the encoder.

*Impact of slice transformer*

To assess the impact of the Transformer in the MST framework, we analyzed the performance when replacing the Transformer with a linear layer to aggregate the information of all slices. Additionally, we tested the performance when averaging the feature vectors. To assess the impact of positional encoding, we compared the performance of MST models with a variant that uses additive positional encoding. Specifically, each slice was assigned a positional embedding based on its slice number, enabling the model to capture spatial dependencies. For all tests, we used DINOv2 as the backbone.

The Transformer model without positional embedding achieved the highest performance across all three datasets (Table 2). Replacing the Transformer with a linear layer or averaging the features resulted in consistently lower AUC values across all datasets.

*Impact of backbone architecture*

To further examine the influence of the feature extractor on MST's performance, we conducted experiments with different backbone architectures. We evaluated the effect of increasing the model size using a larger DINOv2 variant ("base") with 86 million parameters instead of 21 million. We also tested augmenting DINOv2 with register tokens[24] to study if this architectural modification would enhance feature representation and overall performance. Additionally, we assessed the impact of freezing the weights of the DINOv2 backbone during training to determine the necessity of fine-tuning. Lastly, we replaced DINOv2 with a pretrained 2D ResNet to compare self-supervised learning features with those obtained from supervised ImageNet pretraining.

The results are summarized in Table 3. The larger DINOv2 variant performed comparable to the smaller DINOv2. Incorporating registers into DINOv2 did not yield improvements. Freezing DINOv2 weights led to a significant decrease in AUC values on breast MRI and chest CT (reduced to $0.62 \pm 0.03$ and $0.66 \pm 0.03$, respectively), highlighting the necessity of fine-tuning. Replacing DINOv2 with a pretrained ResNet decreased AUC values across all datasets, with a significant drop on the chest CT dataset ($0.92 \pm 0.02$ vs. $0.95 \pm 0.01$; $P = 0.01$).

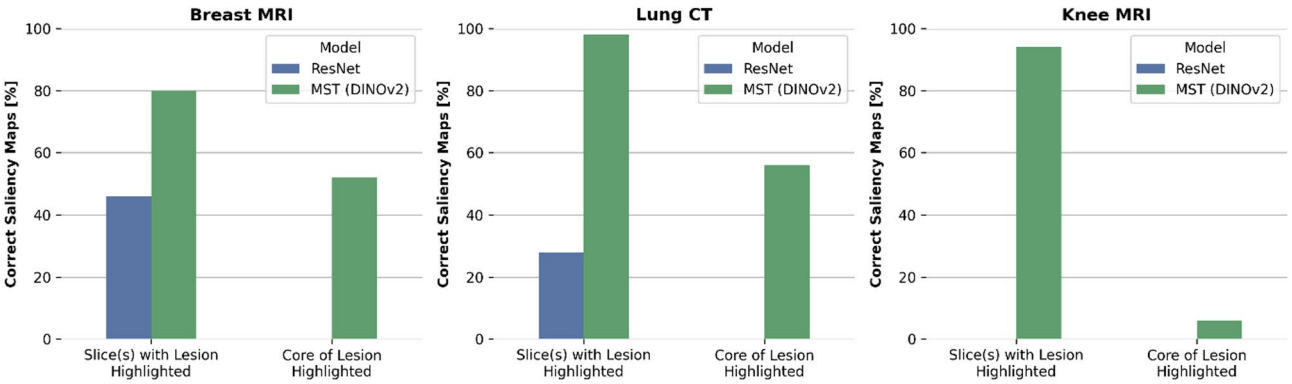



**Fig. 3**. Quantification of saliency map-lesion-correspondence as a function of imaging dataset and model architecture. Percentages of (blinded) radiologist evaluation in terms of slice correctness ("Does the saliency map highlight the slice(s) containing the lesion?"—yes/no) and lesion correctness ("Does the saliency map accurately point to the core of the lesion?"—yes/no).

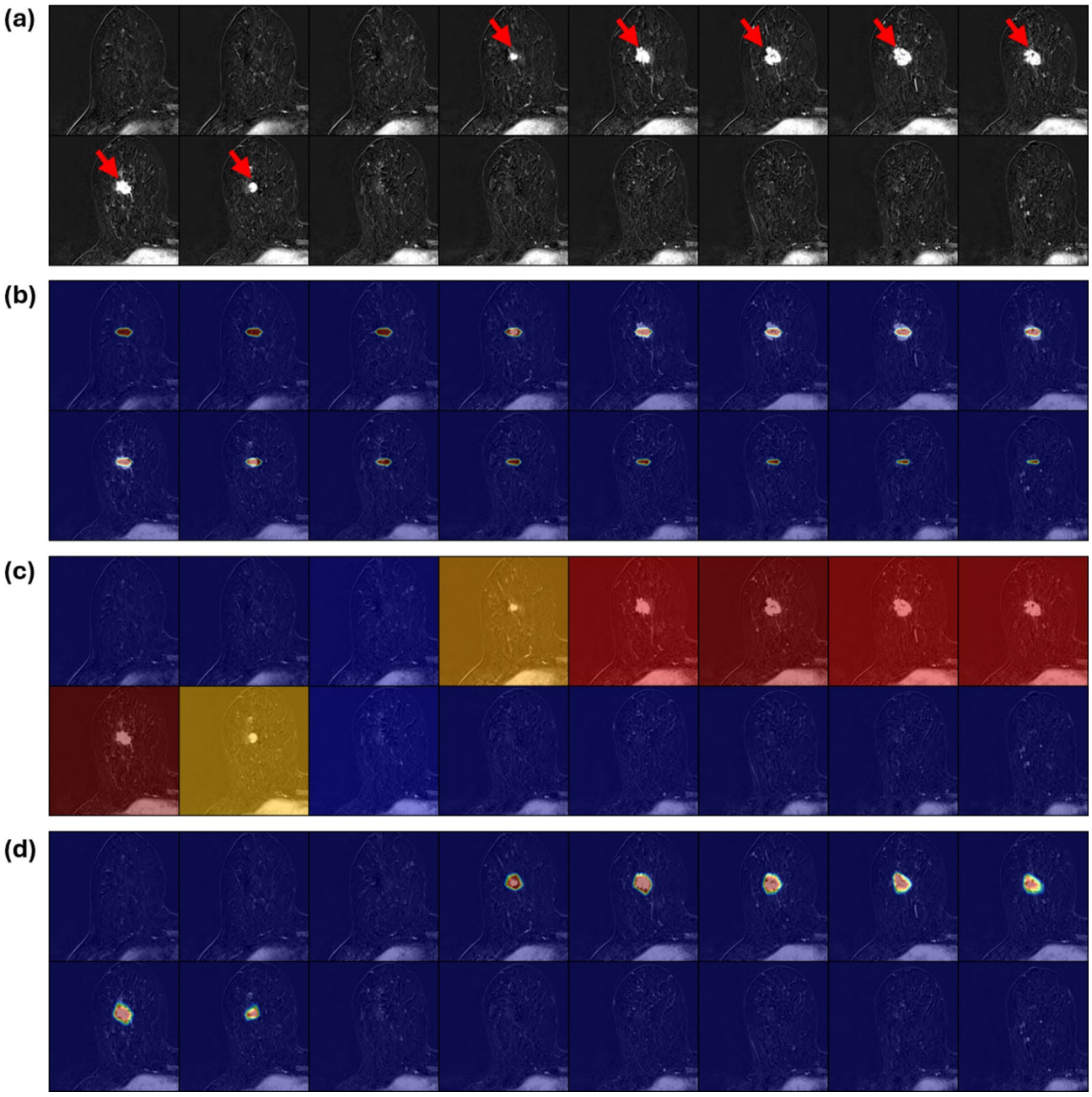



**Fig. 4**. Exemplary saliency maps as a function of model architecture on the breast MRI dataset. (**a**) Consecutive axial slices of an MRI scan of the right breast showing subtraction images with a malignant lesion highlighted by red arrows. (**b**) The saliency maps of the conventional 3D ResNet are spread across the provided image slices and highlight attention on slices without the lesion. The color coding toward blue indicates low attention, while the spectrum toward red indicates high attention. (**c**) In contrast, the MST Slice Transformer focused exclusively on slices with the lesion. (**d**) The combined attention map results from two attention mechanisms, i.e., the Slice Transformer's attention to specific slices and the within-slice attention to specific patches. Thereby, precise localization that corresponds well to the lesion is derived.

The qualitative evaluation showed that the saliency maps of the MST-DINOv2 model were more focused than ResNet or DINOv2-registers as backbones (see Supplementary Figs. S1–S4).

## Discussion

This study addresses two pivotal challenges in applying deep learning to medical imaging: enhancing diagnostic accuracy and improving model explainability through more accurate saliency maps. By introducing the MST framework, which adapts 2D self-supervised models like DINOv2 to 3D medical images, we demonstrate significant advancements in both areas.

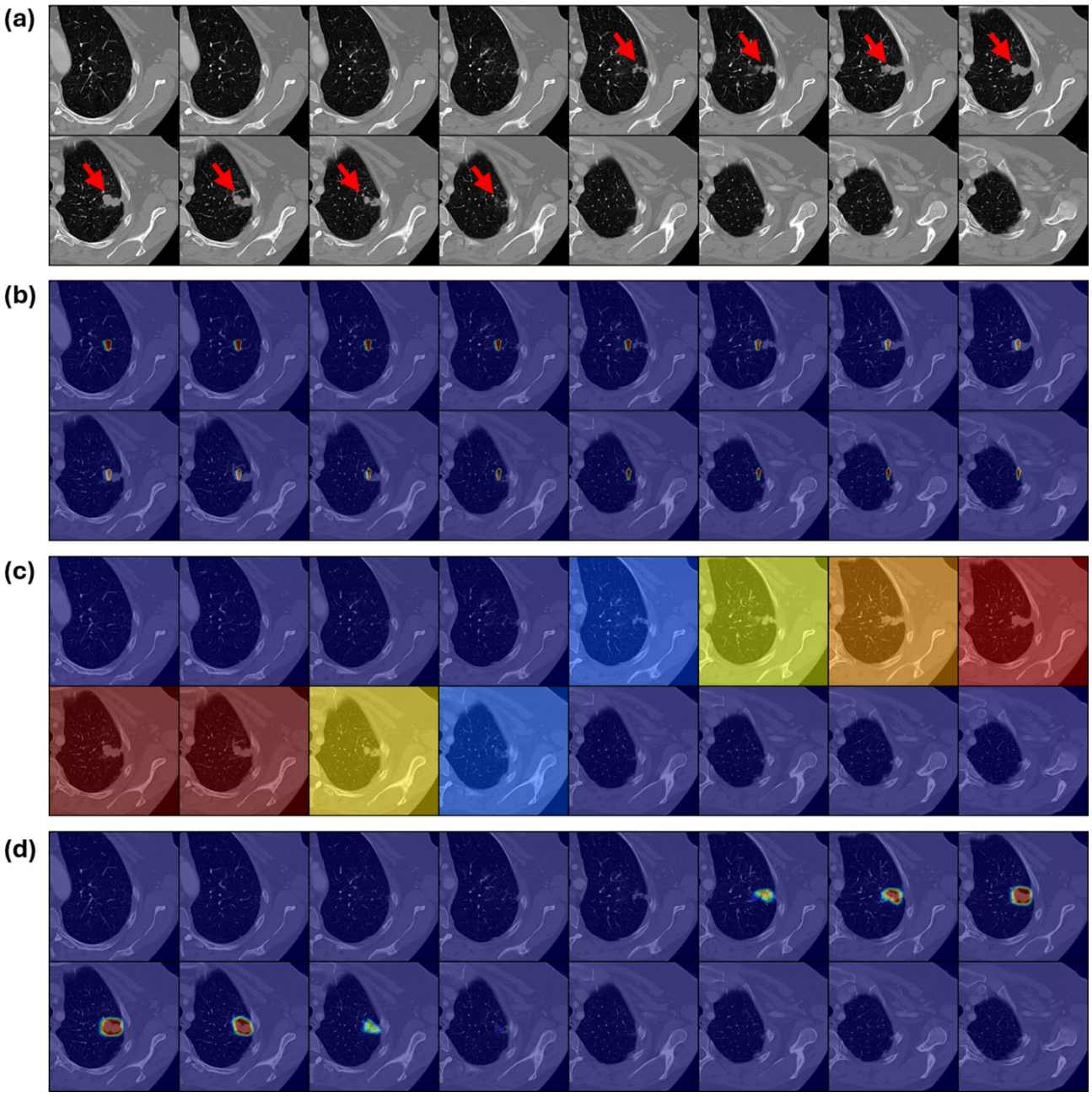



**Fig. 5**. Exemplary saliency maps as a function of model architecture on the lung CT dataset. (**a**) Consecutive axial slices of the left lung centered around a malignant pulmonary nodule are highlighted by red arrows. (**b**–**d**) Image organization as in Fig. 4.

Our MST model achieved superior diagnostic performance across diverse datasets—including breast MRI, chest CT, and knee MRI—evidenced by higher AUC values compared to conventional 3D ResNet architectures. This finding underscores the efficacy of leveraging 2D self-supervised models for 3D medical imaging tasks, extending the benefits observed in previous studies focused on 2D modalities[25–27]. Previous studies have demonstrated the effectiveness of slice-wise feature extraction using Transformer architectures for 3D medical image analysis, primarily in brain MRI[25–27]. However, these efforts did not explore using DINOv2 as a self-supervised feature extraction backbone. Our study demonstrates the effectiveness of DINOv2 across a range of modalities, including breast MRI, knee MRI, and chest CT.

At least of equal importance is the substantial improvement in model explainability. The attention mechanisms inherent in the Transformer architecture provided more precise and informative saliency maps than traditional Grad-CAM methods used with convolutional neural networks. This enhanced localization of pathologies aligns with findings from prior research indicating that Transformer-based methods yield more accurate visual explanations in medical imaging[19,28,29]. This improvement is important since improved saliency maps aid in

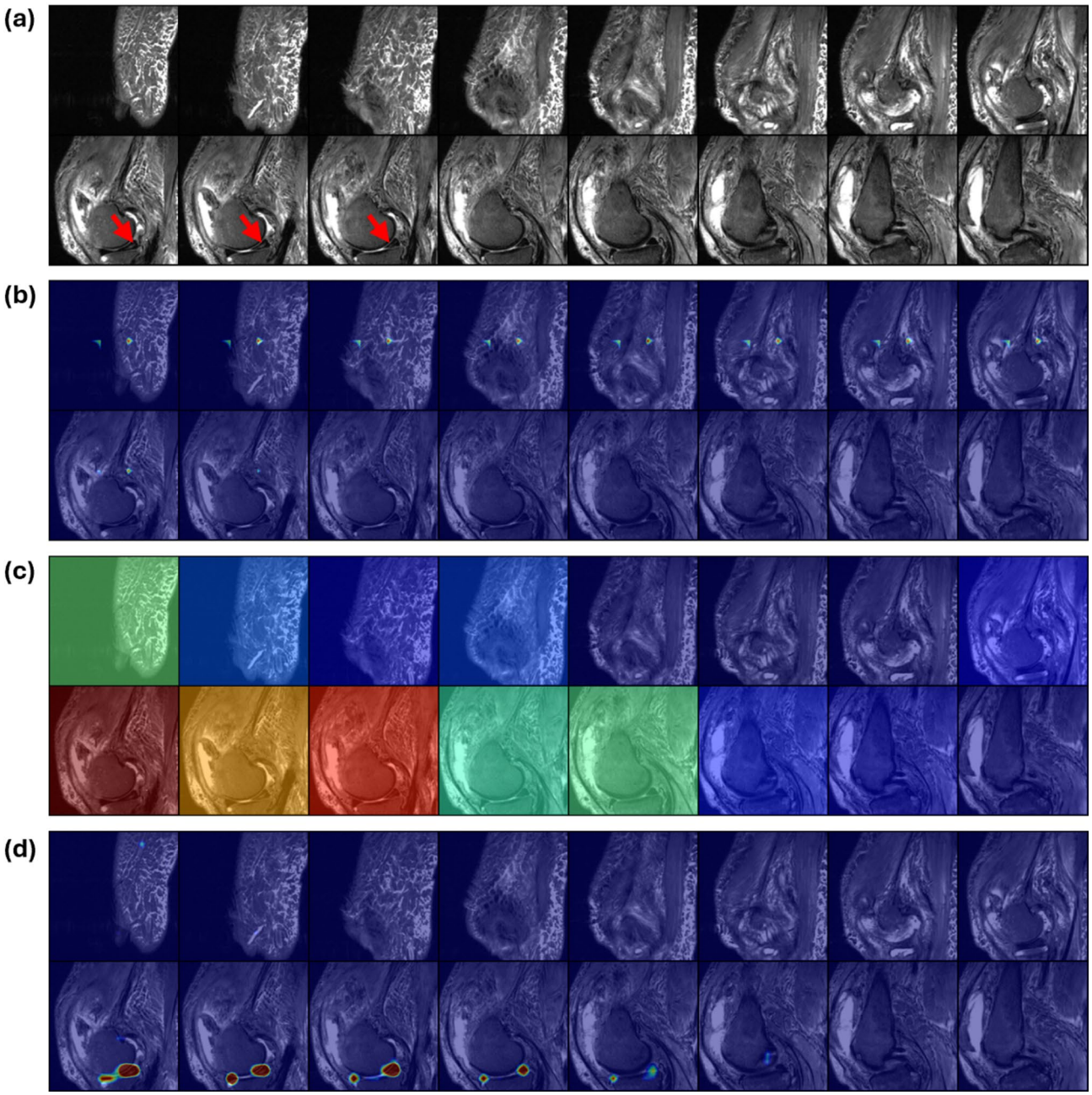



**Fig. 6**. Exemplary saliency maps as a function of model architecture on the knee MRI dataset. (**a**) Consecutive sagittal slices of an MRI study (T2-weighted with fat saturation) of a knee with a medial meniscus tear (posterior horn) highlighted by red arrows. (**b**–**d**) Image organization as in Fig. 4.

understanding the model's decision-making process and increase the trustworthiness of deep learning models in clinical settings[5–7].

However, limitations need to be acknowledged. The evaluation of localization performance using "Slice Correctness" and "Lesion Correctness" is based on heuristics and is inherently limited by subjectivity. While metrics such as Intersection over Union would offer a more objective quantification of overlap between model-predicted regions and ground truth annotations, precise segmentation of all lesions was not feasible within the scope of this study. Similarly, while involving multiple blinded radiologists and assessing inter-rater reliability would enhance the robustness of the evaluation, this was beyond the resources available for this work. Furthermore, the computational demands of processing high-resolution 3D imaging data pose practical challenges: CT scans can have hundreds of slices and spatial resolutions of $512 \times 512$ pixels or higher, exceeding current GPU memory capacities. This constraint necessitates downsampling, which may result in the loss of critical diagnostic information. The current study focused on single MRI sequences (i.e., subtraction images and T2-weighted fat-saturated sequences) and unenhanced CT images for this proof-of-concept in medical imaging

| Dataset | Method | Slice correctness | Lesion correctness |
|---|---|---|---|
| Breast | 3D ResNet | 23/50 | 0/50 |
| Lung | 3D ResNet | 14/50 | 0/50 |
| Knee | 3D ResNet | 0/50 | 0/50 |
| Breast | MST | 40/50 | 26/50 |
| Lung | MST | 49/50 | 28/50 |
| Knee | MST | 47/50 | 3/50 |

**Table 1**. Performance of 3D ResNet- and MST-derived saliency maps in highlighting slices with lesions and cores of lesions as a function of dataset. "Slice correctness" refers to the model correctly highlighting the slice(s) containing the lesion. "Lesion correctness" refers to the model accurately pointing to the core of the lesion.

| | Breast MRI | Chest CT | Knee MRI |
|---|---|---|---|
| Transformer | **0.94 ± 0.01** | **0.95 ± 0.01** | **0.85 ± 0.04** |
| Transformer (with AdPE) | 0.94 ± 0.01 ($P = 0.34$) | 0.93 ± 0.02 ($P = 0.17$) | 0.83 ± 0.04 ($P = 0.41$) |
| Linear layer | 0.88 ± 0.02 ($P < 0.001$) | 0.93 ± 0.02 ($P = 0.26$) | 0.78 ± 0.04 ($P = 0.07$) |
| Average | 0.91 ± 0.02 ($P = 0.03$) | 0.92 ± 0.02 ($P = 0.07$) | 0.82 ± 0.04 ($P = 0.31$) |

**Table 2**. Classification performance as a function of slice aggregation and dataset. Results are shown as mean accuracy ± standard deviation. The best-performing model for each dataset is highlighted in bold. P-values are given for paired comparisons to the transformer. *AdPE* additive positional Embedding.

| | Breast MRI | Chest CT | Knee MRI |
|---|---|---|---|
| MST (DINOv2) | **0.94 ± 0.01** | **0.95 ± 0.01** | **0.85 ± 0.04** |
| MST (DINOv2 base) | 0.93 ± 0.02 ($P = 0.69$) | 0.93 ± 0.02 ($P = 0.27$) | 0.83 ± 0.04 ($P = 0.41$) |
| MST (DINOv2 registers) | 0.93 ± 0.02 ($P = 0.41$) | 0.92 ± 0.02 ($P = 0.02$) | 0.82 ± 0.04 ($P = 0.31$) |
| MST (DINOv2 frozen) | 0.62 ± 0.03 ($P < 0.001$) | 0.66 ± 0.03 ($P < 0.001$) | 0.79 ± 0.04 ($P = 0.14$) |
| MST (ResNet) | 0.93 ± 0.01 ($P = 0.55$) | 0.92 ± 0.02 ($P = 0.01$) | 0.78 ± 0.04 ($P = 0.07$) |

**Table 3**. Classification performance as a function of the transformer model's backbone variant and dataset. Results are shown as mean accuracy ± standard deviation. The best-performing model for each dataset is highlighted in bold.

analysis. In the clinic, multiple contrast phases and reconstructions are used for CT evaluations, and multiple sequences and orientations are used for MRI. Current technical limitations that narrow the input data stream (i.e., the number of input image datasets) may limit the model's ability to fully capture the complex anatomic and pathologic information. Future research is needed to explore the integration of different imaging modalities to enhance model performance and robustness.

In conclusion, adapting 2D self-supervised models like DINOv2 to 3D medical imaging enhances diagnostic accuracy and significantly improves model explainability through better saliency maps. These advancements may provide a cornerstone for overcoming key barriers in deploying deep learning models in clinical practice. Future research should focus on refining these methods and integrating multi-sequence and multi-phase cross-sectional imaging data as well as multimodal data to enhance diagnostic capabilities further.

## Methods

### Dataset collection and preprocessing

We used three publicly available datasets to train and test our approach (Table 4).

The DUKE dataset comprises patients with unilateral or bilateral breast cancer acquired at Duke University Medical Center, USA, between January 2000 and March 2014[20,30]. The following sequences and orientations were available for each MRI study: a non-fat saturated, axial T1-weighted sequence and an axial T1-weighted fat-saturated gradient echo sequence (pre-contrast), and typically four post-contrast, axial T1-weighted fat-saturated sequences. Of these, we used the axial T1-weighted pre-contrast and first post-contrast sequences. Each MRI study was separated into left and right breast image stacks, with each side labeled as either cancerous or non-cancerous (i.e., containing a malignant breast lesion or not). We excluded 271 MRI studies where the specific cancer laterality was not provided. Consequently, 651 MRI studies of 651 patients (mean age, 53 ± 12 years [± SD]) were included. These studies were resampled to a resolution of $0.7 \times 0.7 \times 3.0$ mm$^2$ to standardize voxel dimensions. Subtraction images were generated by subtracting the first post-contrast images from the pre-contrast images to enhance contrast differences indicative of malignancy versus benign tissue changes. Finally,

| Dataset name (reference) | Modality | Anatomic region | Patient number [*n*] (pre-exclusion) | Patient number [*n*] (post-exclusion) |
|---|---|---|---|---|
| DUKE[20] | MRI | Breast | 922 | 651 |
| LIDC-IDRI[21] | CT | Chest | 1010 | 722 |
| MRNet[22] | MRI | Knee | 1199 | 1199 |

**Table 4**. Included datasets and patient cohorts. Data included breast MRI studies from the Duke-Breast-Cancer-MRI (DUKE) dataset, chest CT studies from the lung image database consortium and image database resource initiative (LIDC-IDRI) dataset, and knee MRI studies from the MRNet dataset.

the images were cropped to $256 \times 256 \times 32$ voxels, centered based on foreground-background contrast using a 90% threshold. During training, a $224 \times 224 \times 32$ crop was applied at a randomly selected location, whereas during inference, a center crop was used.

The LIDC-IDRI dataset contains chest CT scans from 1010 patients with 7371 lung nodules[21]. The dataset was acquired by seven academic centers in the United States and eight medical imaging companies using various CT scanners to facilitate the development, training, and evaluation of computer-aided diagnosis systems for lung cancer detection. The CT scans were acquired with slice thicknesses of 0.6–5.0 mm (mean 1.9 mm), in-plane pixel sizes of 0.46–0.98 mm (mean 0.67 mm), and a resolution of $512 \times 512$ voxels with 65–764 slices (mean 240). For 928 lung nodules (of 7371), segmentation outlines were provided by all four expert radiologists. We used the pylidc library[31] to aggregate the segmentation outlines by combining annotations based on consensus among the radiologists. More specifically, ground-truth segmentations were established on a per-voxel basis based on a consensus of at least two (of four) expert radiologists (i.e., $\geq 50\%$ consensus). Practically, the radiologists had delineated the lung nodules and rated the likelihood of malignancy on a 5-point scale ranging from "1: Highly Unlikely" to "5: Highly Suspicious." We excluded 135 patients with nodules smaller than 3 mm because these small nodules are clinically insignificant, challenging to accurately delineate, and may introduce noise. Additionally, 18 nodules were excluded due to their proximity, making them unresolvable by pylidc. The average malignancy rating was calculated for the remaining annotations and on a per-nodule basis. 1026 nodules with an average rating of three were excluded because these intermediate ratings represent uncertain cases, which could introduce ambiguity. Nodules with an average rating of four and five were classified as malignant, while those with an average rating of one or two were classified as benign. The chest CT scans were then cropped to a spatial size of $224 \times 224 \times 32$ voxels centered around the nodule segmentation outline. Altogether, chest CT images containing 1625 lung nodules (1123 benign, 502 malignant) from 722 patients, with an average of two lung nodules/patient, were included.

The MRNet dataset was curated by Stanford University to further the development of machine learning for knee MRI[22]. The publicly available dataset (without the private test set) comprises MRI studies from 1199 patients collected between January 2001 and December 2012 using 1.5 T and 3T MRI scanners. It includes 1250 knee MRI scans with coronal T1-weighted, sagittal T2-weighted fat-saturated, and axial proton density-weighted fat-saturated sequences. Of these, 449 MRI scans (36%) had meniscal tears. The sagittal T2-weighted fat-saturated sequences were chosen for primary analysis, as they are considered particularly relevant for evaluating meniscal pathology[32,33]. The majority vote of three expert MSK radiologists with 6–19 years (average 12 years) of clinical experience was used as the reference label to indicate whether the medial or lateral meniscus was intact or torn. Intact was defined as normal, degenerated without a tear, or post-surgically altered without a tear. Torn was defined as increased signal intensity reaching the articular surface on at least two adjacent slices or a morphologically altered tissue configuration. To standardize the input data for our deep learning model, the images were cropped at a consistent size of $150 \times 150 \times 32$ pixels and resampled to $224 \times 224 \times 32$ voxels.

## Model implementation and training

For the breast MRI and chest CT datasets, we divided the samples into training-validation and test sets with an 80–20% split, stratified by label distribution. For the knee MRI dataset, the publicly available validation set (consisting of 120 MRI studies) was used as the test set. The training-validation set was subdivided into 80% training and 20% validation.

We employed ResNet50[23] with three-dimensional convolutional kernels as our reference model. The model architecture included an input convolutional block, max pooling, four down-sampling convolutional blocks, average pooling, and a final linear layer. The model was implemented using the MONAI[34] library.

Our proposed Medical Slice Transformer (MST) architecture comprised an encoder using the Transformer[15] architecture and an image encoder (Fig. 1). The Transformer encoder consisted of one layer, with 12 or 16 attention heads, and a feed-forward network with 384, 512, or 768 channels, depending on the image encoder used.

The image encoder was exchangeable in that any model that inputs 2D images and outputs 1D feature vectors can be employed for this task. This study focused on ResNet50, pretrained on ImageNet, and DINOv2[35]. For MST-ResNet, we removed the final linear layer and used the 512-dimensional feature vector obtained after the pooling operation. For DINOv2, we used the 384- or 768-dimensional feature vector of the classification token from the final layer.

The feature vectors from all slices were concatenated with a classification token and fed into the slice Transformer. The output vector of the classification token was then passed through a linear layer to obtain the class logits.

All models were trained using a batch size of 2 and 16-bit precision. We employed weighted sampling to address class imbalances. Furthermore, we used the TorchIO library[36] to apply the following data augmentation techniques to address overfitting: random flip, Gaussian noise, random rotation, and random signal inversion.

The training was terminated if the validation AUC values did not improve over 50 epochs, and the checkpoint with the highest validation AUC value was used for testing. The AdamW[37] optimizer was used with a weight decay of $1e^{-2}$ and learning rates of $1e^{-4}$ for 3D ResNet, $1e^{-5}$ for MST-ResNet, and $1e^{-6}$ for MST-DINOv2. Cross-entropy loss was used as the loss function.

## Saliency maps

We employed two techniques to visualize the regions of interest within the images, i.e., those parts of the images the model focused on when making a decision: Gradient-weighted Class Activation Mapping[12] (GradCAM) for convolutional-based models and scaled-dot product attention maps for transformer-based models[16].

*GradCAM*

For the ResNet, we computed gradients for the rectified linear unit function in the last convolutional layer of the ResNet model. The activation outputs were then multiplied by the corresponding gradients to generate saliency maps.

*Attention map*

For the Transformer-based models, we extracted the scaled-dot product attention from the query-key mechanism in the last layer. We focused on the attention weights of the classification token relative to all other tokens, normalizing the weights so that their sum was one. We then averaged the attention weights across all heads.

For the MST model, we multiplied the attention weights from the Transformer with those provided by the image encoder. We applied linear interpolation to align the attention maps with the input image dimensions.

## Statistical analysis

All statistical analyses were performed using Python v3.10 and conducted by G.M.F. and D.T.

The AUC values were calculated using the scikit-learn module[38]. To evaluate whether AUC values between models were statistically significant, Delong's test[39] was applied using a fast implementation[40]. Confidence intervals for the ROC curves were estimated using the bootstrap method with 1000 iterations.

P-values $< 0.05$ were considered statistically significant for all tests.

## Data availability

All datasets in this study are publicly available. DUKE dataset: https://doi.org/10.7937/TCIA.e3sv-re93; LIDC-IDRI dataset: https://doi.org/10.7937/K9/TCIA.2015.LO9QL9SX; MRNet dataset: https://stanfordmlgroup.github.io/competitions/mrnet/.

Received: 11 February 2025; Accepted: 25 June 2025
Published online: 04 July 2025

## References

1. Lång, K. et al. Artificial intelligence-supported screen reading versus standard double reading in the mammography screening with artificial intelligence trial (MASAI): a clinical safety analysis of a randomised, controlled, non-inferiority, single-blinded, screening accuracy study. *Lancet Oncol.* **24** (8), 936–944 (2023).
2. Kuo, R. Y. L. et al. Artificial intelligence in fracture detection: A systematic review and meta-analysis. *Radiology*. **304** (1), 50–62 (2022).
3. Ching, T. et al. Opportunities and obstacles for deep learning in biology and medicine. *J. R. Soc. Interface*. **15** (141), 20170387 (2018).
4. Altaf, F., Islam, S. M. S., Akhtar, N. & Janjua, N. K. Going deep in medical image analysis: concepts, methods, challenges, and future directions. *IEEE Access*. **7**, 99540–99572 (2019).
5. Singh, A., Sengupta, S. & Lakshminarayanan, V. Explainable deep learning models in medical image analysis. *J. Imaging*. **6** (6), 52 (2020).
6. Gulum, M. A., Trombley, C. M. & Kantardzic, M. A review of explainable deep learning cancer detection models in medical imaging. *Appl. Sci.* **11** (10), 4573 (2021).
7. Dhar, T., Dey, N., Borra, S. & Sherratt, R. S. Challenges of deep learning in medical image analysis—improving explainability and trust. *IEEE Trans. Technol. Soc.* **4** (1), 68–75 (2023).
8. Huix, J. P. et al. Are Natural Domain Foundation Models Useful for Medical Image Classification? In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. Waikoloa, HI, USA: IEEE; [cited 2024 Aug 27]. 7619–7628. https://ieeexplore.ieee.org/document/10483777/ (2024).
9. Truong, T., Mohammadi, S. & Lenga, M. How transferable are self-supervised features in medical image classification tasks? 2021 [cited 2024 Jul 1]; https://arxiv.org/abs/2108.10048
10. Nielsen, M., Wenderoth, L., Sentker, T. & Werner, R. Self-Supervision for medical image classification: State-of-the-Art performance with ~ 100 labeled training samples per class. *Bioengineering*. **10** (8), 895 (2023).
11. Huang, Y. et al. Comparative analysis of imagenet pre-trained deep learning models and DINOv2 in medical imaging classification. arXiv; [cited 2024 Aug 27]. https://arxiv.org/abs/2402.07595 (2024).
12. Selvaraju, R. R. et al. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)* [Internet]. Venice: IEEE; 2017 [cited 2024 Aug 17]. 618–626. http://ieeexplore.ieee.org/document/8237336/
13. Shi, Y., Wang, C., Liu, D., Cai, W. & Cabezas, M. Understanding the role of saliency maps for biomarker research in 3D medical imaging classification. In *2023 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*. Port Macquarie, Australia: IEEE; [cited 2024 Aug 29]. 41–48. https://ieeexplore.ieee.org/document/10410916/ (2023).
14. Wiśniewski, M., Giulivi, L. & Boracchi, G. SE3D: A framework for saliency method evaluation In 3D imaging. arXiv; [cited 2024 Aug 29]. (2024). https://arxiv.org/abs/2405.14584

15. Vaswani, A. et al. Attention is all you need. arXiv:170603762 [cs]. 2017 Dec 5 [cited 2021 Mar 29]; http://arxiv.org/abs/1706.03762.
16. Dosovitskiy, A. et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv:201011929 [cs]. 2021 Jun 3 [cited 2022 Mar 28]. http://arxiv.org/abs/2010.11929.
17. Nishigaki, D. et al. Vision transformer to differentiate between benign and malignant slices in 18F-FDG PET/CT. *Sci. Rep.* **14** (1), 8334 (2024).
18. Park, S. et al. Self-evolving vision transformer for chest X-ray diagnosis through knowledge distillation. *Nat. Commun.* **13** (1), 3848 (2022).
19. Wollek, A. et al. Attention-based saliency maps improve interpretability of pneumothorax classification. *Radiol. Artif. Intell.* **5**(2), e220187 (2023).
20. Saha, A. et al. Dynamic contrast-enhanced magnetic resonance images of breast cancer patients with tumor locations. The Cancer Imaging Archive; [cited 2024 Oct 6]. (2022). https://www.cancerimagingarchive.net/collection/duke-breast-cancer-mri/.
21. Armato, S. G. III et al. Data From LIDC-IDRI [Internet]. The Cancer Imaging Archive; [cited 2024 Oct 6]. (2015). https://www.cancerimagingarchive.net/collection/lidc-idri/.
22. Bien, N. et al. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of MRNet. Saria S, editor. *PLoS Med*. **15**(11), e1002699 (2018).
23. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV, USA: IEEE; 2016 [cited 2023 Jun 4], 770–778. http://ieeexplore.ieee.org/document/7780459/ (2016).
24. Darcet, T., Oquab, M., Mairal, J. & Bojanowski, P. Vision transformers need registers. arXiv; [cited 2024 Aug 16]. (2023). https://arxiv.org/abs/2309.16588.
25. Jun, E., Jeong, S., Heo, D. W. & Suk, H. I. Medical transformer: universal encoder for 3-D brain MRI analysis. *IEEE Trans. Neural Netw. Learn. Syst.* 1–11 (2023).
26. Alp, S. et al. Joint transformer architecture in brain 3D MRI classification: its application in alzheimer's disease classification. *Sci. Rep.* **14** (1), 8996 (2024).
27. Jang, J. & Hwang, D. M3T: three-dimensional Medical image classifier using Multi-plane and Multi-slice Transformer. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. New Orleans, LA, USA: IEEE; 2022 [cited 2024 Sep 21], 20686–20697. https://ieeexplore.ieee.org/document/9878673/.
28. Murphy, Z. R., Venkatesh, K., Sulam, J. & Yi, P. H. Visual Transformers and convolutional neural networks for disease classification on radiographs: A comparison of performance, sample efficiency, and hidden stratification. *Radiol. Artif. Intell.* **4** (6), e220012 (2022).
29. Komorowski, P., Baniecki, H. & Biecek, P. Towards evaluating explanations of vision transformers for medical imaging. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Vancouver, BC, Canada: IEEE; 2023 [cited 2024 Sep 26], 3726–3732. https://ieeexplore.ieee.org/document/10208407/.
30. Saha, A. et al. A machine learning approach to radiogenomics of breast cancer: a study of 922 subjects and 529 DCE-MRI features. *Br. J. Cancer*. **119** (4), 508–516 (2018).
31. Hancock, M. C. & Magnan, J. F. Lung nodule malignancy classification using only radiologist-quantified image features as inputs to statistical learning algorithms: probing the lung image database consortium dataset with two statistical learning methods. *J. Med. Imag.* **3** (4), 044504 (2016).
32. De Smet, A. A. How I diagnose meniscal tears on knee MRI. *Am. J. Roentgenol.* **199** (3), 481–499 (2012).
33. Parkar, A. P. & Adriaensen, M. E. A. P. M. ESR essentials: MRI of the knee—practice recommendations by ESSR. *Eur. Radiol.* **34** (10), 6590–6599 (2024).
34. Cardoso, M. J. et al. MONAI: An open-source framework for deep learning in healthcare. [cited 2023 Jul 13]; (2022). https://arxiv.org/abs/2211.02701.
35. Oquab, M. et al. DINOv2: Learning Robust Visual Features without Supervision. arXiv; [cited 2024 Aug 16]. (2023). https://arxiv.org/abs/2304.07193.
36. Pérez-García, F., Sparks, R. & Ourselin, S. TorchIO: A Python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. [cited 2023 Apr 13]; (2020). https://arxiv.org/abs/2003.04696.
37. Loshchilov, I., Hutter, F. & Decoupled Weight Decay, R. 2017 [cited 2024 Jan 10]. https://arxiv.org/abs/1711.05101.
38. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
39. DeLong, E. R., DeLong, D. M. & Clarke-Pearson, D. L. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*. **44** (3), 837 (1988).
40. Sun, X. & Xu, W. Fast implementation of delong's algorithm for comparing the areas under correlated receiver operating characteristic curves. *IEEE Signal. Process. Lett.* **21** (11), 1389–1393 (2014).


## Acknowledgements

The data used in this publication was managed using the research data management platform Coscine with storage space granted by the Research Data Storage (RDS) of the DFG and Ministry of Culture and Science of the State of North Rhine-Westphalia (DFG: INST222/1261-1 and MKW: 214-4.06.05.08-139057).
The authors gratefully acknowledge the computing time provided to them at the NHR Center NHR4CES at RWTH Aachen University (project number p0021834). This is funded by the Federal Ministry of Education and Research, and the state governments participating on the basis of the resolutions of the GWK for national high performance computing at universities (www.nhr-verein.de/unsere-partner). The authors acknowledge the National Cancer Institute and the Foundation for the National Institutes of Health, and their critical role in the creation of the free publicly available LIDC/IDRI Database used in this study.

## Author contributions

G.M.-F. wrote the main manuscript. G.M.-F., S.N., and D.T. conceived the experiments. R.S. analyzed the saliency maps. All authors reviewed the manuscript.

## Funding

Open Access funding enabled and organized by Projekt DEAL.

## Declarations

## Competing interests

DT received honoraria for lectures by Bayer, GE, and Philips and holds shares in StratifAI GmbH, Germany

and in Synagen GmbH, Germany.FK holds shares in StratifAI GmbHJNK declares consulting services for Bioptimus, France; Owkin, France; DoMore Diagnostics, Norway; Panakeia, UK; AstraZeneca, UK; Mindpeak, Germany; and MultiplexDx, Slovakia. Furthermore, he holds shares in StratifAI GmbH, Germany, Synagen GmbH, Germany, and has received a research grant by GSK, and has received honoraria by AstraZeneca, Bayer, Daiichi Sankyo, Janssen, Merck, MSD, BMS, Roche, Pfizer and Fresenius.Other authors have no competing interest.

## Ethics declarations

This study was carried out in accordance with the Declaration of Helsinki.

## Additional information

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-09041-8.

**Correspondence** and requests for materials should be addressed to G.M.-F.

**Reprints and permissions information** is available at www.nature.com/reprints.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2025