nature biomedical engineering

Article

https://doi.org/10.1038/s41551-024-01257-9

Accurate prediction of disease-risk factors from volumetric medical scans by a deep vision model pre-trained with 2D scans

Received: 12 June 2023

Accepted: 23 August 2024

Published online: 01 October 2024

Check for updates

A list of authors and their affiliations appears at the end of the paper

The application of machine learning to tasks involving volumetric biomedical imaging is constrained by the limited availability of annotated datasets of three-dimensional (3D) scans for model training. Here we report a deep-learning model pre-trained on 2D scans (for which annotated data are relatively abundant) that accurately predicts disease-risk factors from 3D medical-scan modalities. The model, which we named SLIViT (for 'slice integration by vision transformer'), preprocesses a given volumetric scan into 2D images, extracts their feature map and integrates it into a single prediction. We evaluated the model in eight different learning tasks, including classification and regression for six datasets involving four volumetric imaging modalities (computed tomography, magnetic resonance imaging, optical coherence tomography and ultrasound). SLIViT consistently outperformed domain-specific state-of-the-art models and was typically as accurate as clinical specialists who had spent considerable time manually annotating the analysed scans. Automating diagnosis tasks involving volumetric scans may save valuable clinician hours, reduce data acquisition costs and duration, and help expedite medical research and clinical applications.

Biomedical imaging analysis is a critical component of clinical care, with widespread use across multiple domains. For example, analysing optical coherence tomography (OCT) images of the retina allows ophthalmologists to diagnose and follow up on ocular diseases, such as age-related macular degeneration (AMD), and tailor appropriate and personalized interventions to delay the progression of retinal atrophy and irreversible vision loss[1,2]. Another example is the analysis of heart function using cardiac imaging, such as heart computed tomography (CT) and ultrasound. Monitoring heart function can help cardiologists assess potential cardiac issues, prescribe medications to improve a medical condition, such as reduced heart ejection fraction, and guide treatment decisions[3,4]. Lastly, radiologists' analysis and regular monitoring of breast imaging such as mammography and magnetic resonance imaging (MRI) help detect early breast cancers, initiate a consequent interventive therapy and determine the effectiveness of such therapeutics[5,6]. These medical insights and actionable information are obtained following an expert's time-intensive manual analysis. The automation of these analyses using artificial intelligence may further improve healthcare as it reduces costs and treatment burden[7].

Deep vision models, such as convolutional neural networks (CNNs) and their derivatives, are considered state-of-the-art methods to tackle computer vision tasks in general[8,9] and biomedical-related vision tasks in particular[10]. To train a deep vision model to accurately learn and predict a target variable in a general vision task (excluding segmentation tasks) from scratch, a very large number of training samples are needed. Transfer learning addresses this challenge by pre-training a vision model for a general learning task on a very large dataset and then using this general model as a starting point for training a specialized model on a much smaller dataset[11]. The key advantage of transfer learning is that the pre-training can be done on a large dataset in another domain, where data are abundant, and then the fine-tuning can be done using a small dataset in the domain of interest. This approach is

✉e-mail: orenavram@gmail.com; ssadda@doheny.org; ehalperin@cs.ucla.edu

specifically useful in clinical environments or when considering emerging biomedical-imaging modalities, where the available data are often very limited. Using a transfer learning approach, a plethora of previously developed deep vision methods analysing two-dimensional (2D) biomedical-imaging data[12–15] were first pre-trained on over a million labelled natural images (in a supervised fashion) taken from ImageNet[16] and, later on, fine-tuned to a specific biomedical-learning task on a much smaller number of labelled biomedical images (typically fewer than 10,000). Some methods used self-supervised-based transfer learning techniques relying mainly on unlabelled biomedical data[17–19], and others combined both natural and biomedical images[7,20]. Overall, the understanding that pre-trained weights can be leveraged as 'prior knowledge' for fine-tuning downstream learning tasks was a core factor in the fruitfulness of the majority of these 2D biomedical-imaging deep vision models.

Many diagnoses rely, however, on volumetric biomedical imaging (for instance, 3D OCT or 3D MRI scans), and transfer learning is not directly applicable, as in contrast to the 2D domain there is no large annotated 'ImageNet-like' dataset of structured 3D scans. Moreover, annotating 3D biomedical images is far more labour prohibitive than 2D images. For example, a 3D OCT scan that is composed of 97 2D frames (usually referred to as B-scans) normally requires a 5–10 min inspection of a highly trained clinical retina specialist to detect retinal-disease biomarkers, such as the volume of a drusen lesion[21]. Therefore, considering the resources typically devoted to such a task, it is practically infeasible to annotate 100,000 (or more) volumes to eliminate the necessity of supervised transfer learning. In fact, even merely compiling such a large-sized volumetric dataset (without labels) that is required for self-supervised-based learning[22] could be cost, processing and storage prohibitive[19] when standard resources are available[23,24]. These gaps are acute because state-of-the-art models for 3D image analysis, such as 3D ResNet[25] and 3D Vision Transformer[26] (ViT), involve the optimization of a very large number of parameters, thus requiring large datasets for training[27].

Nonetheless, several attempts were undertaken to tackle volumetric-biomedical-imaging learning tasks with sparsely annotated training datasets on different data modalities. For instance, SLIVER-net was designed for binary classification of AMD biomarkers in 3D OCT scans[28]. EchoNet was designed to predict heart ejection fraction in ultrasound videos[29]. A few other recent studies achieved state-of-the-art performance using 2D-slice-CNN-based methods and 3D-ResNet-based architectures in diagnosing Alzheimer's disease[30], breast cancer[31] and Parkinson's disease[32] in 3D MRI scans. It is worth noting that, although 3D ResNet was first published in 2018, it is still largely considered a solid baseline and, evidently, very popular not only in MRI studies (for example, in refs. 31,32) but also across other recent volumetric-biomedical-imaging modality studies such as ultrasound[33] and CT[34] studies. The main limitation of each of these approaches is that they are all tailored and optimized for a specific biomedical-imaging modality and domain. While each data modality requires a specific treatment, there are commonalities across the different data modalities, and a foundational approach that can provide improved results across multiple modalities will provide a faster development time for future predictive models. UniMiSS[19], a pioneering pyramid U-like Medical Transformer, has recently been proposed to tackle this gap by using multimodal unlabelled biomedical images in a self-supervised manner. UniMiSS surpassed a diverse set of strong self-supervised approaches[35–39] in a variety of biomedical-imaging learning tasks with different data modalities. However, with respect to volumetric imaging, it was tested on a single classification problem in a single imaging modality (CT) while including this same imaging modality in its pre-training, and regression was not addressed at all. Thus, the full utility of transfer learning across different modalities of volumetric-biomedical-imaging technologies has yet to be attained.

In this Article, we present the slice integration by vision transformer (SLIViT) framework, a uniform 3D-based deep-learning model that overcomes the annotation bottleneck and is adept at volumetric-biomedical-imaging learning tasks. We leverage the combination of a 2D ConvNeXt-based[40] feature-map extractor and a tweaked ViT[41] together with cross-dimension and cross-domain (that is, imaging modality, organ and pathology) transfer learning. The 2D-based feature-map extractor allows leveraging previous 2D biomedical (and non-biomedical) vision knowledge when extracting information from a given volume in a variety of biomedical-imaging modalities. Then, the attention-based mechanism of the ViT allows integration of the extracted information across the 2D frames (henceforth interchangeably referred to as 'slices') of the volume in question and reconstruction of long-range dependencies of the volume's depth dimension. We demonstrate the generalizability and utility of SLIViT in very different biomedical domains, including retinal-disease risk biomarkers diagnosis in 3D OCT scans, cardiac function in heartbeat ultrasound videos, hepatic-disease severity assessment from 3D MRI liver scans and pulmonary nodule-malignancy screening in 3D CT chest scans. We show that SLIViT consistently attains significantly improved performance compared with both strong generic baselines and domain-specific state-of-the-art models. Notably, SLIViT provides these improved performance results across data modalities with neither tailoring the architecture nor extensively optimizing hyperparameters per (task or) data modality, unlike other biomedical-imaging learning methods (for instance, refs. 7,13,19). We further show that SLIViT is robust to frame permutation, suggesting that it could be applied to datasets in which the slice order (within a volume) is not recorded. Finally, we demonstrate that SLIViT's performance is comparable to clinical specialists' manual annotation and that it shortens the annotation time by a factor of 5,000; hence it can potentially be used to save resources, reduce the burden on clinicians and expedite ongoing research[7].

Results

A unified artificial-intelligence framework for analysing volumetric-biomedical-imaging data

In this study, we devise SLIViT, a deep-learning vision model for automatic annotation of clinical features in 3D biomedical images. An overview of SLIViT is summarized in Fig. 1. SLIViT preprocesses volumes into 2D images, which then pass through two 2D-based deep vision architectures: (1) a ConvNeXt backbone module[40] that extracts a feature map for the slices (that is, 2D frames of a volume) and (2) a ViT module[41] that integrates this feature map into a single diagnosis prediction. One key part of SLIViT is that its feature-map extractor is initialized by pre-trained weights. These weights are obtained by pre-training a ConvNeXt first on ImageNet[16] and then on a 2D biomedical-imaging dataset. This pre-trained network can then be fine-tuned for different diagnostic tasks in volumetric-biomedical-imaging data using a relatively small training set (down to a few hundred samples). The premise behind this approach is that different biomedical-imaging modalities share a common set of visual features, and so, a network trained on one 2D biomedical-imaging learning task could serve as a useful training starting point for another network that deals with a volumetric-biomedical-imaging learning task.

To cope with volumetric data, we treat each volume as a set of slices. A similar technique, also known as 2.5D, was shown to be effective for volumetric data modalities[42]. Essentially, each original slice of the volume is embedded into a single feature map. However, to reduce memory overhead[43], the slices are tiled and processed as a single elongated 2D image (rather than a set of separate slices), such that it conforms with the input dimension expected by the 2D-based feature-map extractor. Once the feature map is extracted, it is divided into patches, each (roughly) representing features extracted from the corresponding original slice. The patches are then paired with (trainable) positional embeddings and comprehensively aggregated

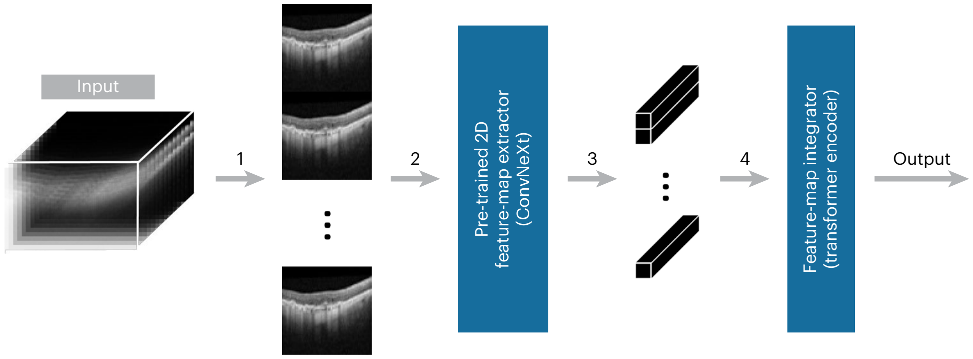

Input
1
2
Pre-trained 2D feature-map extractor (ConvNeXt)
3
4
Feature-map integrator (transformer encoder)
Output

Fig. 1 | The SLIViT framework. The input of SLIViT is a 3D volume of N frames of size $H \times W$. (1) The frames of the volume are resized and vertically tiled into an 'elongated image'. (2) The elongated image is fed into a ConvNeXt-based feature-map extractor that was pre-trained on both natural and biomedical 2D labelled images. (3) An $8N \times 8 \times 768$ (3D) feature map is extracted and partitioned into N patches of size $8 \times 8 \times 768$, each (roughly) representing features extracted from the corresponding original frame. (4) Patches are fed into a ViT-based feature-map integrator followed by a fully connected layer that outputs the prediction for the task in question (see Methods for further details).

using a downstream ViT module[41]. SLIViT's ViT module together with (trainable) positional embeddings allows preserving of the long-range dependencies across the depth dimension if needed[30,44]. Similar divide-and-conquer schemes were shown to be fruitful in other studies as well[25,42,45,46]. It is worth noting that the fact that the ViT considers dependencies between frames in the feature space implicitly eliminates the necessity for image registration preprocessing.

We pre-trained SLIViT with a 2D OCT B-scan dataset[47] and tested it on six datasets of four different volumetric-biomedical-imaging data modalities (OCT, ultrasound, MRI and CT) with a limited number of annotated samples, tackling a variety of clinical-feature learning tasks (including both classification and regression). We evaluated the diagnosis performance of several ocular disease high-risk factors[28] (OCT) and malignant pulmonary nodules (CT) and measured it by both the receiver operating characteristic (ROC) area under the curve (AUC) and precision-recall (PR) AUC. In the ultrasound and MRI experiments, we compared the R^2 of the models' predictions versus ground truth in (respectively) cardiac function analysis and hepatic fat level imputation. In each data modality, we compared SLIViT with a diverse set of up to six strong baselines, including domain-specific[19,25,28–30] and generic (fully-supervised-based[25,26] and self-supervised-based[7,19]) state-of-the-art methods. SLIViT manifested consistent and significant performance superiority across domains (Fig. 2). In the following sections, we present these and additional results in detail.

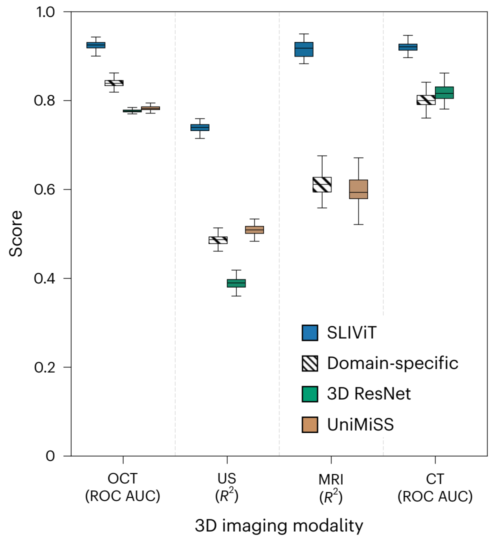


Fig. 2 | Overview of SLIViT's performance across 3D imaging modalities. The performance scores in four different volumetric-biomedical-imaging learning tasks: eye-disease biomarker diagnosis in 3D OCT scans (classification), heart-function analysis in ultrasound (US) videos (regression), liver fat level imputation in volumetric MRI scans (regression) and lung malignant cell-aggregation screening in 3D CT scans (classification). The domain-specific methods (hatched) used are SLIVER-net, EchoNet, 3D ResNet and UniMiSS for OCT, ultrasound, MRI and CT, respectively. The cross-modality benchmarking used are 3D ResNet and UniMiSS, which are (fully) supervised-based and self-supervised-based, respectively (see relevant experiment's section for additional benchmarking). The expected R^2 and ROC AUC of a random model are 0 and 0.5, respectively. Box plot whiskers represent a 90% CI.

Detecting ocular disease high-risk factors using 3D OCT scans

We first compared SLIViT's performance against trained SLIVER-net, 3D ResNet, 3D ViT and UniMiSS models on the Houston Dataset which includes only 691 OCT volumes of different individuals (Methods). OCT volume data were collected from independent individuals affected in at least one eye by AMD, a globally leading cause of irreversible central visual impairment[48]. Each OCT volume had four different binary labels of AMD high-risk biomarkers[49] procured by a senior retina specialist—drusen volume larger than 0.03 mm^3 (DV), intraretinal hyperreflective foci (IHRF), subretinal drusen deposits (SDD) and hyporeflective drusen cores (hDC). We randomly split the dataset into train, validation and test sets of sizes 483 (70%), 104 (15%) and 104 (15%), respectively, and trained four different SLIViT models (one per binary label). We used both ROC AUC and PR AUC scores (the latter is also known as average precision or average positive predictive value) for performance evaluation. The models were trained (using less than 600 volumes) and tested on the same split (left panels of Fig. 3 and Extended Data Fig. 1, and Supplementary Table 1). In all four biomarkers, SLIViT significantly outperformed the other approaches in both evaluation metrics. For example, in the DV classification task (also shown as the OCT experiment in Fig. 2) SLIViT (ROC AUC = 0.924; confidence interval (CI) [0.909, 0.938]) was significantly better compared with the second-best performing method (SLIVER-net ROC AUC = 0.838; CI [0.813, 0.86]; paired *t*-test $P < 0.001$). In terms of average precision of the DV classification, SLIViT (PR AUC = 0.914; CI [0.898, 0.928]) significantly outperformed the second-best performing method (3D ResNet PR AUC = 0.759; CI [0.748, 0.769]; paired *t*-test $P < 0.001$). It is worth noting that, as the biomarkers considered in these experiments are all structural, their identification requires the aggregation of 3D information. Thus, the ability of SLIViT to successfully identify these biomarkers suggests that it adequately captures a 3D signal within a given volume.

To further challenge SLIViT, we sought to explore its performance on the SLIVER-net Dataset used in the original SLIVER-net study[28]. In this task, SLIVER-net should have an advantage as it was optimized for

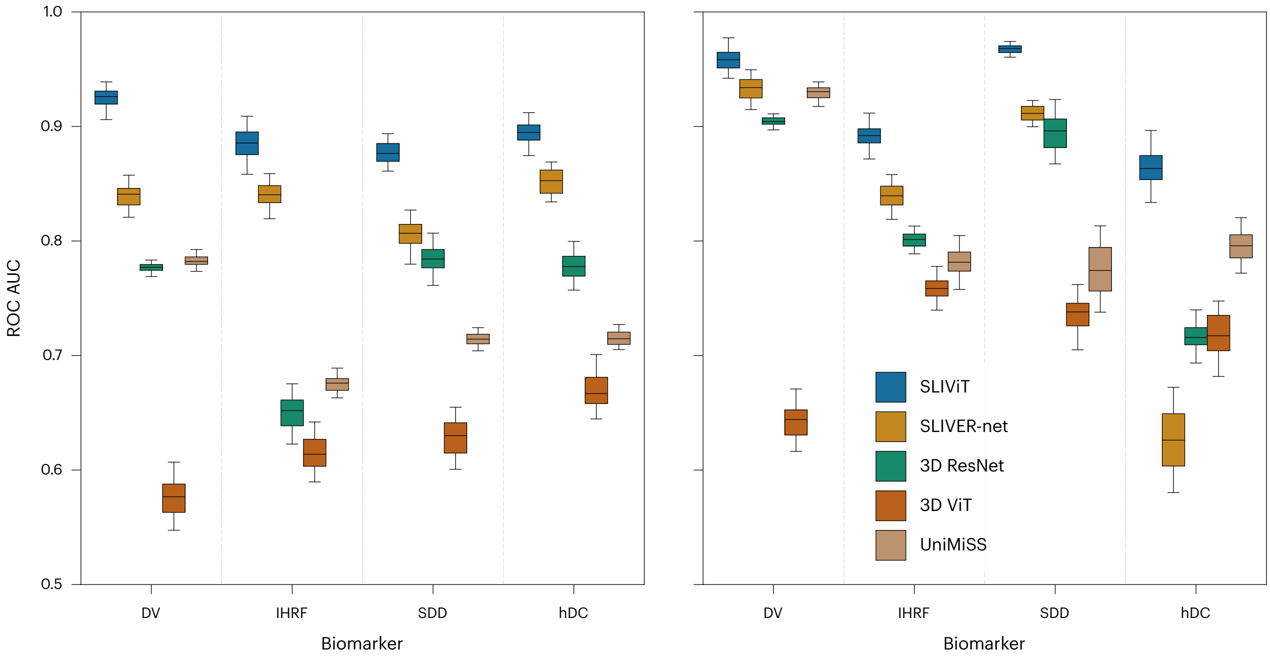


Fig. 3 | Performance comparison on four tasks of AMD-biomarker classification when trained on less than 700 OCT volumes. The ROC AUC scores of SLIViT, SLIVER-net, 3D ResNet, 3D ViT and UniMiSS on four binary classification problems of AMD high-risk factors (DV, IHRF, SDD and hDC) in two independent 3D OCT datasets. Left: the performance when trained and tested on the Houston Dataset (Supplementary Table 1). Right: the performance when trained on the Houston Dataset and tested on the SLIVER-net Dataset (Supplementary Table 2). The expected performance of a random model is 0.5. Box plot whiskers represent a 90% CI.

this dataset. The SLIVER-net Dataset was composed of roughly 1,000 OCT scans collected from three different clinical centres (Methods). We trained SLIViT, SLIVER-net, 3D ResNet, 3D ViT and UniMiSS, this time using all the 691 Houston Dataset volumes, and used the SLIVER-net Dataset as the test set. For some biomarker classification tasks, the relative improvement of SLIViT compared with SLIVER-net was reduced, as expected in this setting. Yet, SLIViT was never overperformed by the other approaches, in any of the four AMD-biomarker classification tasks (right panels of Fig. 3 and Extended Data Fig. 1, and Supplementary Table 2).

Analysing cardiac function using heartbeat ultrasound videos

To evaluate SLIViT's generalizability, we next tested it on other 3D data modalities. The EchoNet-Dynamic Dataset contains 10,030 standard apical four-chamber view ultrasound videos (echocardiograms) obtained from unrelated individuals. Each echocardiogram was labelled with a continuous number representing the corresponding ejection fraction measured in a clinical setting[50]. The ejection fraction is a key metric of cardiac function as it measures how well the heart's left ventricle is pumping blood. Low ejection fraction measurements (<0.5) can indicate cardiomyopathy or other heart problems[3,51]. As a first experiment, we sought to explore SLIViT's ability to predict cardiomyopathy as a binary classification task. To this end, we binarized the ejection fraction measurements accordingly (≥0.5 was considered as normal[52,53]) and, using the original EchoNet-Dynamic Dataset split, trained SLIViT and 3D ResNet (Supplementary Fig. 1). SLIViT obtained 0.913 ROC AUC (CI [0.901, 0.928]) and significantly overperformed 3D ResNet with 0.793 ROC AUC (CI [0.772, 0.814]) (paired t-test $P < 0.001$).

In a second experiment, we sought to test SLIViT in a regression task. EchoNet, a GoogLeNet-based architecture, was previously developed for predicting the ejection fraction of a given echocardiogram and obtained a 0.5 R^2 on the EchoNet-Dynamic Dataset test set[29]. This reported result did not include a CI (that would allow a direct comparison), and the trained model itself was not published. Thus, we implemented the proposed method and were able to reproduce similar levels of performance ($R^2 = 0.489$; CI [0.434, 0.526]). Using the same split from the original EchoNet paper, we then trained SLIViT and obtained a significant improvement of 0.75 R^2 (CI [0.706, 0.781]; paired t-test $P < 0.001$). As we did in all other experiments, we also tested 3D ResNet and UniMiSS and observed that both significantly underperformed SLIViT with 0.384 (CI [0.364, 0.413]) and 0.502 (CI [0.487, 0.531]) R^2, respectively (ultrasound experiment in Fig. 2, and Fig. 4). A scatter plot of the actual-versus-predicted per trained model is shown in Supplementary Fig. 2. Moreover, we also examined (1) a factorized spatiotemporal ResNet architecture (R(2 + 1)D, in contrast to the 3D-filter-based R3D ResNet we used across our study; 'Benchmark specifications') that is known to capture well both spatial and temporal features from video frames and achieved state-of-the-art performance in a variety of video-based learning tasks[25], and (2) 3D ViT[26]. Both methods performed below par compared with the other above-mentioned benchmarks ($R^2 = -0.081$; CI [−0.106, −0.056] and $R^2 = 0.333$; CI [0.27, 0.396], respectively).

This result, together with the exceptional magnitude of this public homogenous volumetric biomedical dataset, further motivated us to examine the dynamics of the training set size and SLIViT's performance in predicting the ejection fraction of a given echocardiogram (Fig. 4). We randomly sampled size-decreasing subsets from the original training set and trained a SLIViT model per subset. Compared with other examined methods trained on the original training set ($n = 7{,}465$), when SLIViT used the 25% subset ($n = 1{,}866$), its performance ($R^2 = 0.487$; CI [0.466, 0.507]) was significantly better than R3D, R(2 + 1)D and 3D ViT (paired t-test $P < 0.001$); on par with EchoNet (paired t-test $P > 0.579$); and significantly lower than UniMiSS (paired t-test $P < 0.001$). When SLIViT used the 50% subset, it significantly outperformed all other benchmarked methods ($R^2 = 0.614$; CI [0.594, 0.634]; paired t-test $P < 0.001$). These observations substantiate SLIViT's ability to appropriately learn spatiotemporal features using a sparsely labelled dataset.

Predicting hepatic fat levels in 3D MRI liver scans

We next sought to evaluate SLIViT's ability to model 3D MRI data. We used the United Kingdom Biobank (UKBB) Dataset containing 3D

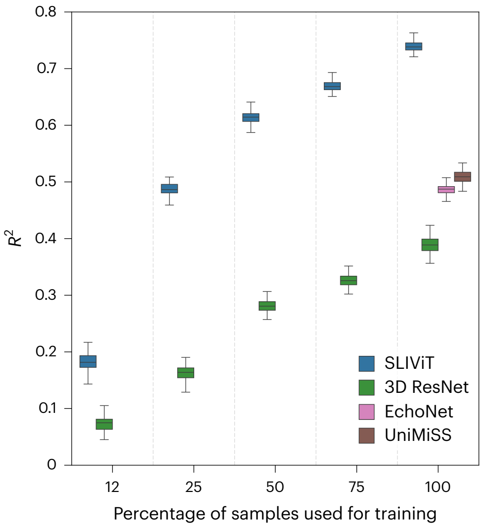


Fig. 4 | Performance comparison on cardiac function prediction tasks when trained on echocardiograms. The R^2 scores of SLIViT, 3D ResNet, EchoNet and UniMiSS on heart ejection fraction prediction. Several SLIViT models were trained, each on a different-sized training subset (sampled from the original training set). The *x* axis shows the sampled subset size (in percentage) used for training, where 100% corresponds to the original training set. Box plot whiskers represent a 90% CI. It is worth noting that SLIViT, when trained on 25% ($n = 1{,}866$) of the original training set, obtained similar accuracy as the other examined methods trained on 100% ($n = 7{,}465$) of the original training set.

hepatic MRI scans and a corresponding measurement for hepatic proton density fat fraction (PDFF) level. The PDFF measurement provides an accurate estimation of hepatic fat levels, and it is also proposed to be used as a non-invasive method to limit unnecessary hepatic biopsies[54–56]. The development of a quantitative measurement of fat has been instrumental in improving the diagnosis of various fatty-liver and diabetes-related diseases[57–61]. We removed unlabelled scans and preprocessed the rest of the dataset to contain only a single scan per individual. In this experiment we compared SLIViT with 3D ResNet (which achieved state-of-the-art performance in a variety of recent MRI-related artificial-intelligence-based studies[30–32]) and UniMiSS. We randomly split the dataset and trained both models to measure PDFF levels of a given 3D MRI. SLIViT reached 0.916 R^2 (CI [0.879, 0.952]) and significantly outperformed both 3D ResNet and UniMiSS that obtained 0.611 (CI [0.566, 0.644]) and 0.599 (CI [0.531, 0.667]) R^2, respectively (paired *t*-test $P < 0.001$; MRI experiment in Fig. 2). We also evaluated the performance of 3D ViT and a recently developed 2D-slice-CNN-based architecture that was shown to perform well on volumetric MRI learning tasks[30], but they both ended up with poor performance compared with all the above-mentioned benchmarks ($R^2 = 0.18$ (CI [0.145, 0.214]) and −0.130 (CI [−0.111, −0.148]), respectively).

Classifying nodule malignancy in 3D CT chest scans

To further demonstrate SLIViT's cross-modality generalizability, we evaluated it on 3D CT data. To this end, we used the NoduleMNIST3D Dataset, containing 3D thoracic CT scans, each (binary) labelled for nodule malignancy[62]. In the United States, more than a million patients are diagnosed with pulmonary nodules each year, and these nodules are observed in roughly 30% of thoracic CT scans. As in other biomedical-imaging domains, the scan screening is subjective and depends on the clinical specialist's experience (for example, small nodules may be missed[63]). Efficient and accurate assessment could hasten malignant pulmonary nodule treatment and reduce unnecessary testing when benign[64]. Using the dataset's predefined split, we trained SLIViT, 3D ResNet and UniMiSS and compared the results on the test set (CT experiment in Fig. 2). It is worth noting that UniMiSS was pre-trained on 3D thoracic CT scans and, thus, has a potential advantage. Yet, SLIViT obtained 0.926 ROC AUC (CI [0.904, 0.947]) and 0.785 PR AUC (CI [0.758, 0.837), significantly overperforming (paired *t*-test $P < 0.001$) both UniMiSS with 0.8 ROC AUC (CI [0.765, 0.836]) and 0.627 PR AUC (CI [0.568, 0.685]), and 3D ResNet with 0.821 ROC AUC (CI [0.776, 0.857]) and 0.619 PR AUC (CI [0.508, 0.718]). We also evaluated the performance of 3D ViT that obtained 0.873 ROC AUC (CI [0.825, 0.914]) and 0.713 PR AUC (CI [0.627, 0.792]). Here, as well, SLIViT was significantly superior considering both performance metrics (paired *t*-test $P < 0.001$).

SLIViT efficiently attains the quality of clinical specialists

To showcase the potential utility of automating the detection of AMD high-risk biomarkers, we gathered the Pasadena Dataset, a third 3D OCT dataset containing 205 3D OCT volumes of (205) independent individuals. The ground truth for this dataset was obtained by three masked senior retina specialists (we used a majority vote when there was no consensus). We asked seven masked junior clinicians to annotate each of the OCT volumes in this dataset for the aforementioned four AMD high-risk biomarkers, that is, DV, IHRF, SDD and hDC. We also annotated these volumes using the same SLIViT model we trained on the 691 Houston Dataset volumes. Figure 5 and Extended Data Fig. 2 summarize, respectively, the true positive rate (also known as recall) versus false positive rate (also known as false alarm rate) and the positive predictive value (also known as precision) versus recall of SLIViT and the seven junior clinicians over the Pasadena Dataset. Clinicians typically reached comparable performance but had to invest 5,000-fold more time to do so (on average, it took 17 working hours net for each clinician to procure the annotations, while SLIViT completed the job in under 12 s). It is worth noting that SLIViT obtained considerably lower performance in the hDC classification task compared with the other biomarker classification tasks. A possible reason is the absence of a universal consensus on the clinical definition of hDC. This feature had the highest senior specialists' annotation discordance among the four biomarkers, suggesting indeed that it is harder to distinguish between affected and unaffected individuals.

SLIViT is robust to within-volume frame permutation

We next sought to explore SLIViT's robustness to changes in the order of the frames encoding a volume. To this end, we generated 100 copies of the Houston Dataset and randomly shuffled each volume (in each of these 100 copies). Then, we used the same split to train 100 SLIViT models (one per shuffled copy; henceforth 'shuffled models') and one model on the Houston Dataset using the original order (henceforth 'original model') to classify the aforementioned AMD high-risk factors. Extended Data Fig. 3 shows the average bootstrapped ROC AUC dispersion of these 101 models. It is worth noting that the original model did not outperform the shuffled models. We observed that compared with the 100 shuffled-models performance, the average rank of the original model across the four AMD biomarkers was 40. This finding suggests that even if the original order is not documented, SLIViT's performance does not deteriorate. Thus, not only does SLIViT effectively aggregate information across slices, it can do this even when the order of slices is not maintained.

We found this frame-permutation invariance intriguing, especially when considering the fact that the examined biomarkers are structural. We were thus motivated to examine more deeply the utility of the attention mechanism in our model. We conducted an experiment in which we trained a tweaked SLIViT model with only one multi-head attention layer (instead of five in the original version) in which non-immediately adjacent attention weights are set to 0. This tweaked version can pool information only from immediately adjacent frames in the volume. We trained the tweaked SLIViT model for DV classification on a random split of the (original) Houston Dataset and evaluated the performance on the test set twice—once with frames in order (0.912 ROC AUC

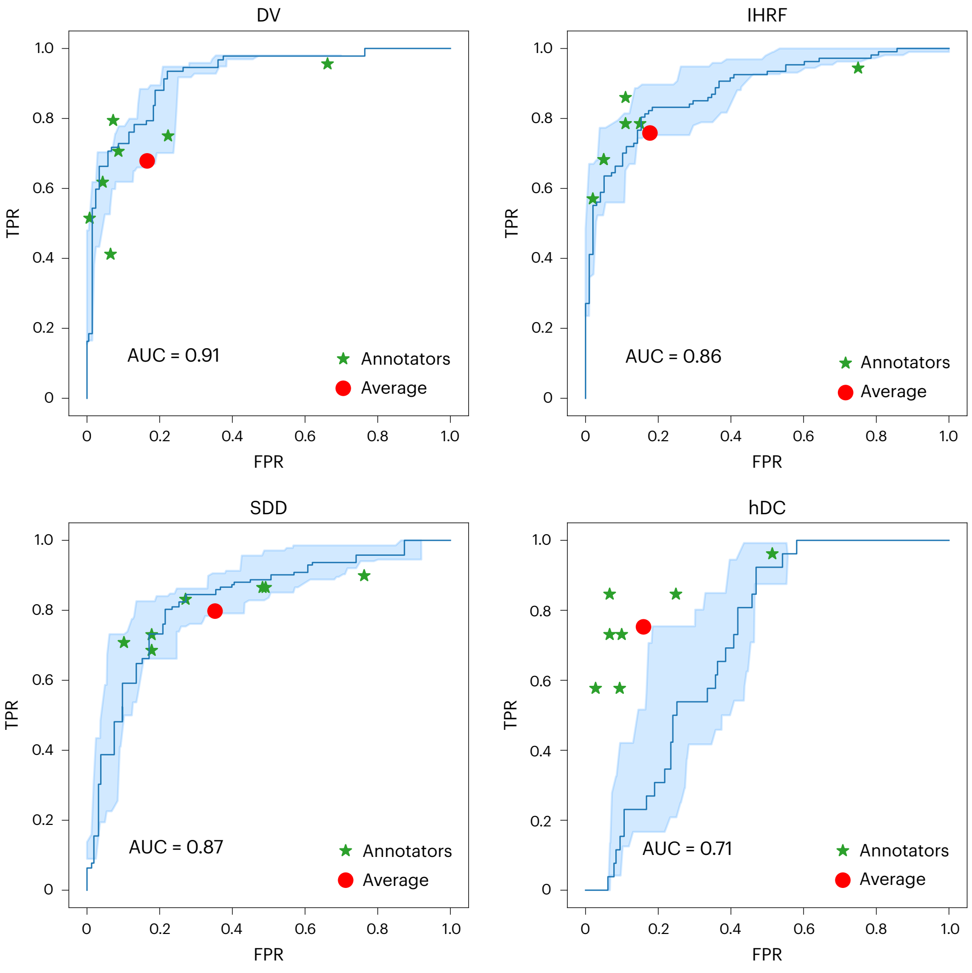


Fig. 5 | SLIViT's performance compared with manual assessment by retina clinical specialists. The ROC curves (blue) of SLIViT trained to predict four AMD high-risk biomarkers (DV, IHRF, SDD and hDC) using less than 700 OCT volumes (Houston Dataset) and tested on an independent dataset (Pasadena Dataset). In each panel, the light-blue shaded area represents a 90% CI for SLIViT's performance, and the red dot represents the clinical specialists' average performance. The green asterisks correspond to the clinical specialists' manual assessments. Two of the clinical specialists obtained the exact same performance score for IHRF classification. TPR, true positive rate; FPR, false positive rate.

(CI [0.91, 0.919]) and 0.908 PR AUC (CI [0.901, 0.914])) and once with random frame shuffling (0.846 ROC AUC (CI [0.832, 0.862]) and 0.827 PR AUC (CI [0.805, 0.838])). The noticeable decline in the performance (0.066 and 0.081 in ROC and PR AUC scores, respectively) of the model when evaluated on the shuffled test set suggests that SLIViT's performance on this learning task relies on successfully pooling information from different frames.

The utility of pre-training SLIViT

The utility of ImageNet pre-training (henceforth 'ImageNet weights') has been demonstrated in various biomedical-imaging learning tasks[7,12,14,15,65–67]. That said, transfer learning between unrelated domains remains fairly controversial[18,29,68–70]. Moreover, commonalities across biomedical-imaging data modalities may be counterintuitive. We thus conducted a comprehensive pre-training ablation study across the different learning tasks to evaluate the benefit of our cross-modality and cross-dimensionality transfer learning approach and assess the contribution of different selections made for the pre-training step of SLIViT (Extended Data Figs. 4–8).

ImageNet pre-training

We first wished to assess the contribution of ImageNet pre-training and thus compared the four different initializations: random weights, ImageNet weights, random weights initialization followed by 2D OCT B-scans pre-training (henceforth 'Kermany weights') and ImageNet weights initialization followed by 2D OCT B-scans pre-training (henceforth 'combined weights', which are the weights used in all experiments described above). The results of this experiment (Extended Data Figs. 4 and 5) indicate three key insights that concord with conclusions indicated in previous studies[7,18,71]. First, we observed that using ImageNet weights improved performance for all the data modalities we tested relative to random weights. We also see that using 2D OCT B-scans in pre-training with either Kermany weights relative to random weights or combined weights relative to ImageNet weights improved performance in all downstream learning tasks. It is worth noting that, in the four OCT-related classification tasks, using Kermany weights (that is, without ImageNet) was the best approach and typically led to better performance, even when compared with the combined approach (Extended Data Fig. 4). Moreover, only pre-training strategies that leveraged the 2D OCT B-scan dataset at full, that is, Kermany weights and combined weights, showed consistent superior performance relative to all other tested benchmark methods (left panels of Fig. 3 and Extended Data Fig. 1, and Extended Data Fig. 4). That being said, in all ultrasound, MRI and CT experiments, SLIViT typically achieved superior performance relative to all benchmark methods tested, regardless of the pre-training strategy (Fig. 2 and Extended Data Fig. 5). This finding shows the advantage of SLIViT's architecture for cross-modality volumetric-biomedical-imaging learning tasks.

Self-supervised pre-training

We next wished to assess the benefit of using supervised learning for pre-training, as opposed to self-supervised learning. The latter was demonstrated as a powerful approach in different visual tasks[72], specifically in the biomedical-imaging domain where procuring annotations is laborious and expensive[7,17,19,20]. We thus sought to explore the utility of self-supervised-based pre-training approach on SLIViT using an unlabelled version of the 2D OCT B-scans dataset. To this end, we took the REMEDIS approach[7]. The performance of REMEDIS was shown to be robust to different self-supervised techniques. We thus followed REMEDIS default scheme and used SimCLR[73] as the self-supervised technique. Nevertheless, any other CNN-based self-supervised approach, such as MoCo[74], RELIC[75] and Barlow Twins[76], could theoretically be leveraged by SLIViT. REMEDIS was originally shown to obtain remarkable performance when pre-trained even on much smaller (unlabelled) datasets than our 2D OCT B-scans dataset. Yet, initializing SLIViT with the fully supervised pre-trained weights significantly outperformed the self-supervised initialization in all downstream learning tasks (paired t-test $P < 0.001$; Extended Data Figs. 4 and 5). It is worth noting that the same performance-superiority conclusion regarding the competitor benchmarks from the previous section held for the self-supervised-based version of SLIViT, implying its potential to harness unlabelled data when available.

2D biomedical-imaging pre-training

We also sought to compare the utility of different 2D biomedical-imaging data modalities. We previously defined Kermany weights that were obtained by random weights initialization followed by 2D OCT (using the Kermany Dataset) pre-training of SLIViT's feature-map extractor backbone. Similarly, we now define 'organ weights' and 'chest weights', obtained by random weights initialization followed by (respectively) 2D CT (using Organ{A,C,S}MNIST[62]) and 2D X-ray (using ChestMNIST[62]) pre-training. We used Kermany weights, organ weights and chest weights to conduct the following two experiments. First, we wished to examine the similarities between the representations learned by biomedical-weights-initialized SLIViT backbones (without downstream-task-specific fine-tuning). To this end, we initialized three backbones with Kermany weights, organ weights and chest weights (henceforth 'biomedical backbones') and two additional baselines, one with ImageNet weights and another with random weights (henceforth 'ImageNet backbone' and 'random backbone', respectively). We then took one of our datasets, projected it through each of these five backbones and measured the centred kernel alignment (CKA) similarity index[77] between pairs of projections (Extended Data Fig. 6). We considered two CKA score distributions for our comparison—the distribution of the top-5% CKA scores (henceforth 'top-5% distribution'), which are likely to be enriched with informative features, and the distribution of the overall CKA scores (henceforth 'overall distribution'). First, we observed that each of the three biomedical backbone projections was more similar to the other two biomedical backbone projections than to the random and the ImageNet backbone projections (t-test $P < 0.001$). This finding held when we compared not only the corresponding top-5% distributions but also the overall distributions. Even when comparing the overall distribution of two biomedical backbone projections with the top-5% distribution of a biomedical backbone and a non-biomedical backbone projections, we observed convincingly robust results. For example, the overall distribution for the Kermany backbone projection and each of the other two biomedical backbone projections was comparable to the top-5% distribution of Kermany and ImageNet backbone projections (t-test $P > 0.05$) and significantly higher compared with the top-5% distribution of Kermany and random backbone projections (t-test $P < 0.001$). The same was observed for the other two biomedical backbone projections. These findings confirm our initial hypothesis that different data modalities share a basic set of features.

In the second experiment, we sought to assess the cross-biomedical-imaging modality of SLIViT. We used four 2D biomedical-imaging datasets—Kermany, OrganMNIST, ChestMNIST and Mixed (a dataset made up of images from all three biomedical datasets)—to pre-train four SLIViT models. Each model was initialized using ImageNet weights and then pre-trained on the respective 2D biomedical-imaging dataset. The Mixed-based SLIViT was pre-trained to classify one out of the total 29 classes included in these three 2D modalities (4 for OCT, 11 for CT and 14 for X-ray). Then for each volumetric-biomedical-imaging learning task, we fine-tuned each of the four pre-trained models. As in the other pre-training experiments, we observed that using 2D OCT data in pre-training typically provides a noticeable advantage in the 3D OCT classification tasks (Extended Data Fig. 7). Furthermore, in all other analysed tasks, SLIViT typically achieved superior performance relative to all competitor benchmarks tested, regardless of the 2D biomedical-imaging dataset used for pre-training (Fig. 2 and Extended Data Fig. 8). This discovery further illustrates SLIViT's cross-modality and cross-dimensionality potencies in 3D biomedical-imaging learning tasks.

Discussion

We devised SLIViT, an artificial-intelligence-based framework that allows the accurate analysis of a wide variety of 3D biomedical-imaging datasets. SLIViT leverages a unique combination of deep vision modules and 'prior knowledge' from the 2D domain. This, in turn, allows it to be adept at 3D biomedical-imaging learning tasks, in which the number of training samples is typically very limited (due to labour-prohibitive and cost-prohibitive procurement) and significantly outperforms domain-specific state-of-the-art models.

To showcase SLIViT's effectiveness and generalizability, we evaluated it over several classification and regression problems in diverse biomedical domains (retinal, cardiac, hepatic and pulmonary) across different 3D biomedical-imaging data modalities (OCT, ultrasound, MRI and CT) against domain-specific[19,25,28–30] and generic (fully-supervised-based[25,26] and self-supervised-based[7,19]) state-of-the-art methods. We started by demonstrating SLIViT's superiority when trained on less than 700 volumes in four independent binary classification learning tasks of retinal-disease risk factors with two independent 3D OCT datasets. Then we showed SLIViT's superiority in two heart function analysis tasks both done with an echocardiogram dataset. We next tested SLIViT on an MRI dataset of 3D liver scans labelled with a corresponding hepatic fat content measurement and, again, observed significant improvement compared with the state of the art. We further exhibited SLIViT's supremacy in pulmonary nodule malignancy screening using a CT dataset of 3D chest scans. We also showed that SLIViT was able to obtain on-par performance to clinical specialists' manual assessment but, rather, almost four orders of magnitude faster compared with the annotation procurement net time required by the specialists. Last, we explored SLIViT's learning ability robustness to randomly permuted volumes. We showed that a scenario of shuffled volumes dataset, a recurring situation in the very limited number of publicly available volumetric datasets, has little to no effect on SLIViT's performance, meaning that SLIViT is not only applicable in such scenarios but also potentially agnostic to the imaging protocol.

To facilitate reproducibility, generalizability and the likelihood that other researchers will be able to successfully apply SLIViT to their datasets, we intentionally avoided complex hyperparameter tuning and the usage of specialized hardware for training as required by other methods (as in ref. 19, for example). The thrifty sizes of the different architectures we used were set according to our available (standard) computational resources, and other hyperparameters were set to default values. This suggests that there is room for further improvement in task-specific performance. Yet, in its current form, SLIViT can serve as a reliable foundation model for any study of volumetric biomedical imaging. We believe that SLIViT's simplicity is one of its major strengths.

The utility of self-supervised pre-training has been validated in numerous biomedical-imaging learning tasks[7,19,20,71,78,79]; however, its general translatability across domains remains unclear[22]. According to our study, where a large-enough 2D labelled dataset is accessible and limited labelled volumes are available, the supervised pre-training approach is superior. This finding was supported by our experiments for fine-tuning both in the same domain and across domains. That being said, as demonstrated, SLIViT's pre-training strategy is very flexible and can thus harness the utility of self-supervised approaches, such as REMEDIS or a masked autoencoder[80,81]. If one has access to an(other) large unlabelled dataset of biomedical images (whether 2D or 3D), then self-supervised pre-training SLIViT either as an alternative to or followed/preceded by supervised pre-training using 2D biomedical images may further improve the model's performance. It is worth noting that the end-to-end fine-tuning approach SLIViT takes (Methods) was shown to attain typically better performance for self-supervised-based biomedical-imaging learning tasks[22]. That is, SLIViT already uses an optimized fine-tuning approach for a potential self-supervised-based avenue.

SLIViT was tested on 3D OCT scans, ultrasound videos, 3D MRI volumes and 3D CT images, and can potentially be leveraged to analyse other types of volumetric-biomedical-imaging data modalities, such as 3D X-ray. Such imaging data are inherently structured in the sense that they involve a limited assortment of objects and movements (typically shrinkage, dilation and shivering). SLIViT is specifically tailored to be adept at analysing a series of biomedical frames created in a structured biomedical-imaging process and does not pretend to be proficient at learning problems of natural videos, such as action recognition tasks. Natural videos are inherently more complex, as the background may change and objects may flip, change colour (due to shade) and even disappear (due to obfuscation), let alone when considering a multi-scene video. In addition, there is a plethora of gigantic natural video datasets that allow standard 3D-based vision models to be adequately tuned for natural video learning tasks. We thus do not expect SLIViT to outperform (as is) standard 3D-based vision models in natural-video-learning tasks (such as action recognition). That being said, SLIViT could potentially be tweaked to perform well on natural videos as well, for instance, using a different feature-map extractor; however, this direction requires further research.

There are multiple additional steps that are required to deploy SLIViT in a clinical setting. It is worth noting that the point of operation (trade-off between precision and recall) is application specific, and further optimization may be required to obtain optimal results at that point of operation. We note that point of operation varies also across clinicians (Fig. 5 and Extended Data Fig. 2). Moreover, additional evaluations of the models are required to ensure no systematic biases exist that would lead to increasing health disparities[82,83].

Outlook

This study highlights a substantial step towards fully automating volumetric-biomedical-imaging annotation. The major leap happens under 'real life' settings of a low-number training dataset. SLIViT thrives given just hundreds of training samples for some tasks giving it an extreme advantage over other 3D-based methods, in almost every practical case that is related to 3D biomedical-imaging annotation. Even under the unrealistic assumption that the financial resources are endless, in ongoing research, due to its nature, the hurdle of a limited-size training dataset is inevitable. Once a previously unknown disease-related risk factor is found and characterized, it could take months to train a specialist to be able to accurately annotate this recently discovered risk factor in biomedical images at scale. However, using a relatively small training dataset (that can be annotated within only a few working days of a single trained clinician), SLIViT could dramatically expedite the annotation process of many other non-annotated volumes with an on-par performance level of a clinical specialist. Thus, it may not only reduce the duration and costs associated with data acquisition and save precious clinical specialists' time but also expedite medical research and other clinical applications.

Methods

SLIViT's development and analysis

SLIViT was implemented in Python 3.8 using PyTorch[84] v1.10.2, fast.ai[85] v2.6.3 and scikit-learn[86] v1.0.2 libraries (full libraries and version list can be found at the project's GitHub repository; 'Code availability'). Weights & Biases (https://www.wandb.com/) was used for experiment tracking and visualizations of the training procedures.

Model specifications

The SLIViT framework contains a preprocessing step, a 2D ConvNeXt that serves as a feature-map extractor and a 2D ViT that serves as a feature-map integrator (Fig. 1). The ConvNeXt architecture has several complexities[40]. Here we used the backbone of the tiny variant (ConvNeXt-T) with 256 × 256 image size as SLIViT's feature-map extractor. We conducted an ablation study to evaluate different combinations for the feature-map extractor and feature-map integrator (Supplementary Fig. 3). Using more complex ConvNeXt variants did not lead to performance improvement. Thus, we followed best practices for model generalization and used the simplest variant (ConvNeXt-T) as a feature extractor. The ViT-based feature-map integrator underwent a few adjustments with respect to the original architecture[41], including using Gaussian error linear unit as the activation function[87] and initializing the positional embeddings as the number of the original slice. It is worth noting that we intentionally avoided complex hyperparameter tuning and usage of specialized hardware as required by other methods[19]. The ViT's depth (number of layers, 5) was set according to our available (standard) computational resources to facilitate reproducibility, generalizability and the likelihood that other researchers will be able to successfully apply it to their datasets. The ViT's width is governed by the number of 2D frames of the input volume. That being said, we examined the performance of versions of SLIViT that use different ViT configurations and found that none of them resulted in noticeable improvements compared with the default configuration (Supplementary Fig. 4).

Let N be the number of $H \times W$ 2D frames of an input image, where (H, W) is the resolution of the original frame(s). Given an input $W \times H \times N$ image, its N frames are resized and tiled into an image of size $256N \times 256$ (step 1 in Fig. 1). The preprocessed image is then fed into the feature-map extractor which generates, in turn, an $8N \times 8 \times 768$ feature map. This 3D feature map is partitioned into N different $8 \times 8 \times 768$ 3D 'patches' (corresponding to the terminology used in the original ViT paper[41]). Note that due to the convolution's locality property, each of these patches roughly corresponds to features obtained from a different frame. Each of the N patches is then flattened into a 1D vector (of length $8 \times 8 \times 768$) and then tokenized into a vector of size 768 using a fully connected layer. The patch number (that essentially corresponds to an original slice number) is then added to each of the tokenized patches, and the results are then fed into the ViT (along with a class token of the same size). The ViT outputs N encoded values and a class token. The class token is then fed into another fully connected layer to generate the final output. Using the 2D ViT as a feature-map integrator corresponds with the Factorized Encoder with 'late fusion of depth information' of the previously devised 3D ViT named ViViT[26] yet is far less complex than the 3D ViT. It is worth noting that, while the dependencies across frames are modelled merely at the feature-map level (and thus could be somewhat approximate, in contrast to fully 3D neural networks, which theoretically model dependencies across slices in any defined region), this approach has two main advantages. First, it could eliminate the necessity for image registration preprocessing. Moreover, the feature-space-based aggregation across frames allows SLIViT to be more effectively fitted for small training sets, as shown by its improved performance relative to fully 3D architectures (such as 3D ResNet and 3D ViT).

Pre-training

We borrowed an ImageNet-1K pre-trained (SLIViT-like) feature-map extractor architecture, that is, a ConvNeXt-T backbone, from https://huggingface.co/facebook/convnext-tiny-224, and appended to it a subsequent fully connected layer that fit the number of categories in the pre-training classification task. We then trained this SLIViT-backbone-like module on a publicly available 2D biomedical-imaging labelled dataset. Training the feature-map extractor for 10 epochs on a dataset containing (at least) 100,000 images took less than 3 h using only a single NVIDIA Tesla T4 (16 GB) GPU. Several sets of pre-trained weights were examined in this study ('The utility of pre-training SLIViT'). The pre-trained backbone weights obtained from combining ImageNet initialization with additional pre-training on the Kermany Dataset (henceforth 'combined weights'), which typically led to the best performance, are available in the project's GitHub repository ('Code availability').

Per-task fine-tuning

Each of the SLIViT models used in the different experiments reported here was initialized with the combined weights. The fine-tuning was done in an end-to-end fashion[22]. Namely, rather than merely training the downstream feature-map integrator while keeping the feature-map extractor frozen, all the model's parameters were set as trainable and were then fine-tuned (according to the dataset and task in question). We intentionally avoided complex hyperparameter tuning as required by some other methods (for example, in ref. 19) to facilitate reproducibility and generalizability. Frames were resized into 256 × 256 pixels to fit SLIViT's pre-trained backbone architecture, and then standard preprocessing transformations were applied (including contrast stretching, random horizontal flipping and random resize cropping) using PyTorch's default values. Binary cross entropy and L1 norm were used as loss functions for the classification and regression tasks, respectively. In each experiment, excluding the ultrasound and CT experiments (in which the splits were given), a random validation set was used for determining the convergence of the training process. The model was optimized using the default fast.ai optimizer with the default parameters. The starting learning rate in each training procedure was chosen by fast.ai's learning rate finder, and the model was fitted using the fit-one-cycle approach for faster convergence[88,89]. All models were trained with four samples per batch, and early stopping was set to five epochs, meaning that the training process continued until no improvement was observed in the validation loss for five consecutive passes on the whole training set. The model weights that achieved the lowest loss on the validation set during training were used for the test set evaluation.

Feature similarity analysis

To demonstrate the visual similarity across biomedical domains, we used the CKA similarity index[77]. CKA allows measuring the similarity between the features extracted using any two neural-network layers for a given sample set. Here we sought to compare the output of the feature-map extractor (when initialized with different sets of weights), as it functions as the input for SLIViT's feature-map integrator backbone. We considered five of the backbone versions, each initialized with a different set of pre-trained weights. The sets included three 2D biomedical-imaging-based weights (that were obtained by pre-training on the Kermany[47], Organ{A,C,S}MNIST[62] and ChestMNIST[62] datasets), ImageNet weights and random weights. The CKA similarity scores were computed by projecting the volumes from the NoduleMNIST3D Dataset[62] onto the feature space of each of the five models. The dataset was chosen arbitrarily among our volumetric datasets. For each of the projection pairs (excluding ImageNet-random pair) we computed the CKA scores between corresponding-slice projections of a given volume and averaged the results across slices and volumes (Extended Data Fig. 6). We considered two CKA distributions for our comparison—the top-5%-scores distribution and the overall-scores distribution. The difference significance between the CKA of different pairs of models was assessed using a standard t-test $(H_A : \mu \neq 0)$.

Statistical analyses

The performance of each trained model was evaluated (on the corresponding test set) using an appropriate metric score. The classification tasks were evaluated using the area under the ROC and PR curves. The regression tasks were evaluated using the R^2 metric. The test set predictions were calculated, and a 90% CI was computed for each evaluated score using a standard bootstrapping procedure with 1,000 iterations as done in other studies[17,90]. Briefly, let n denote the test set size; for each bootstrap iteration n samples were randomly drawn (with repetition), and based on the predictions of the sampled set a single score was obtained. Out of the 1,000 sampled-sets score distribution, the 50th and 950th ranked scores were selected to obtain the 90% CI. To compute the significance value of the difference between two given distributions (induced by two different models), a paired t-test on the distribution of differences between the sampled-set corresponding scores was computed $(H_A : \mu \neq 0)$. In any of the t-tests conducted in this study, a difference was considered to be significant if the test produced a P value lower than 0.001 subject to Bonferroni correction for multiple hypothesis testing.

Benchmark specifications

In this study we used several baselines to benchmark SLIViT across the different data modalities. The baselines included SLIVER-net[28], two different types of 3D ResNet[25] (R3D (unless stated otherwise) and R(2 + 1)D), 3D ViT[26], UniMiSS[19], EchoNet[29] and a 2D-slice-CNN-based architecture[30]. As SLIViT, all models were trained with the fit-one-cycle learning-rate scheduler[88,89]. SLIVER-net was subjected to the same pre-training approach as SLIViT, namely, an ImageNet weights initialization followed by supervised pre-training on the Kermany Dataset. EchoNet performance was reproduced (on the same dataset) as in the original paper[29]. The (ImageNet-initialized) 2D-slice-CNN-based were already optimized for MRI-based learning tasks and thus used as is. As for UniMiSS, we used the pre-trained MiT-22 variant that was shown to be best performative across all the tasks examined in the original UniMiSS paper[19]. Although other studies use benchmarks as is while optimizing their method (for instance, refs. 19,28), we did conduct several hyperparameter tuning experiments for the more generic methods we examined (that is, 3D ResNet and 3D ViT) to confirm that their default configurations are reasonable. The experiments complied with the following best practices. Given a dataset, the original validation and test sets were set aside. The original training set was split into sub-train, sub-validation and sub-test, using the same proportions in the original split. The different examined hyperparameter configurations were evaluated using this (sub-)split. For each configuration, the weights of the model with the lowest sub-validation loss were used to evaluate the performance on the sub-test set. It is worth noting that, due to the heavy computational cost and limited resources, only a restricted number of hyperparameter configurations were examined, and we considered only the classification task in the 3D CT dataset. The hyperparameters examined for 3D ResNet were number of layers (18 and 50) and pre-training strategy (random weights and Kinetics400[91] weights). The hyperparameters examined for 3D ViT were width (96 and 192) and depth (4 and 6). Neither configuration evaluation ended up with significant improvements (on the sub-test set), and given the many heavy computations required for this study, we thus preferred the simplest (previously optimized) configuration from each original paper. That is, randomly initialized 18-layer 3D ResNet and factorized spatiotemporal encoder 3D ViT (with default hyperparameters) as it was shown in the original paper to be the best performative variant[26].

In addition, we examined the performance of a self-supervised-based SLIViT using the REMEDIS approach[7]. To this end, we initialized SLIViT with ImageNet weights and then pre-trained it on an unlabelled version of the Kermany Dataset, using SimCLR[73] (REMEDIS' default learning scheme) as the self-supervised strategy (and then fine-tuned it according to the downstream learning task).

The Houston Dataset

At the Retina Consultants of Texas Eye Clinics, 1,128 patients were diagnosed with intermediate AMD in their scanned eye by clinical examination (Beckman Classification[92]) between October 2016 and October 2020. This study was reviewed and approved by the Ethics Committee of Retina Consultants Texas (Houston Methodist Hospital, Pro00020661:1 'Retrospective Prospective Analysis of Retinal Diseases'). As the data collection was retrospective, a waiver of informed consent was granted. In case both eyes of a given patient were eligible, one eye was randomly included in the dataset. The dataset included Heidelberg Spectralis (HRA+Optical Coherence Tomography OCT SPECTRALIS; Heidelberg Engineering) 6 × 6 mm (fovea centred, 10 × 10°; 49 B-scans spaced 122 µm apart, Automatic Real-Time (ART) function = 6) OCT volumes. The data were transferred to the Doheny Image Reading Research Laboratory (DIRRL) for imaging analysis and annotation of the structural OCT biomarkers for AMD progression[93,94]. The AMD-biomarker analysis was conducted at the DIRRL in compliance with the Declaration of Helsinki and approved by the University of California at Los Angeles (UCLA) Institutional Review Board (IRB; Ocular Imaging Study, Doheny Eye Center UCLA). Cases with evidence of late stage of AMD and/or additional macular diseases or poor-quality imaging were excluded from the analysis. In total, 691 eyes (of 691 patients) were eligible for the biomarkers analysis. The annotations were procured by a senior clinical retina specialist. The recorded case frequency in the whole dataset was as follows: (1) 48.34% of the scans had drusen volume >0.03 mm^3 within the central 3 mm^2 (denoted DV); (2) 36.18% of the scans had IHRF; (3) 31.4% of the scans had SDDs; and (4) 11.29% of the scans had hDC. It is worth noting that some scans were positive for more than one biomarker. The positive-label frequencies of the test set were 37.5%, 41.35%, 48.08% and 31.73%, respectively.

The SLIVER-net Dataset

The SLIVER-net Dataset, which was originally used to tune and validate SLIVER-net[28], was collected from three independent medical centres between February 2013 and July 2016 (ref. 95). The dataset consisted of 1,007 OCT volumes; each had 97 B-scans (97,679 B-scans overall) collected from 649 participants of the Amish general population, who had a record of at least one individual with AMD in the family history. Imaging was conducted at three clinical centres in Pennsylvania, Indiana and Ohio under the supervision of investigators at the University of Pennsylvania, University of Miami and Case Western Reserve University, respectively. The research was approved by each of the IRBs of the respective institutions, and all participants signed written informed consent. All OCT volumes in this dataset were acquired with the Heidelberg Spectralis OCT using a scan pattern centred on the fovea (20° × 20°; 97 B-scans; 512 A-scans per B-scan; ART = 9). To fit the Houston Dataset trained model, we down-sampled each of the SLIVER-net Dataset volumes by taking every other B-scan, thus squeezing each volume into 49 B-scans. Also, to avoid aliasing, we applied an anti-aliasing filter on OCT volumes. The positive-label frequencies in this dataset were 3.38%, 7.85%, 1.99% and 2.68% for DV, IHRF, SDD and hDC, respectively. Although the annotations for this dataset included the eyes laterality, the scans themselves lacked the laterality obscuring the link between a scan to its annotation in case both eyes were scanned for a patient. To address this gap, we considered the middle slice per volume to determine the laterality and trained a standard CNN on the Houston Dataset (that had the eyes laterality recorded). Using the trained network (97% accuracy on an external test set; not shown), we inferred the laterality for the SLIVER-net Dataset scans when needed, that is, when both eyes of the same patient were scanned.

The Pasadena Dataset

The Pasadena Dataset established for this study contained 205 OCT volumes (fovea centred, 10 × 10°, ART = 5) collected from 205 individuals at the Doheny Eye Center UCLA in Pasadena between 2013 and 2022. This study was reviewed and approved by the IRB of the UCLA (IRB number 15-000083). Informed consent was waived for study participants given the retrospective nature of the study. Each of the OCT volumes was acquired on the Heidelberg Spectralis HRA+Optical Coherence Tomography (OCT SPECTRALIS; Heidelberg Engineering). Out of the 205 OCT volumes, 198 contained 97 B-scans and 7 contained 49 B-scans. The OCT volumes were independently annotated by ten DIRRL-certified clinical retina specialists: three seniors (expert retina specialists) and seven juniors. The ground truth for this dataset was determined by the senior retina specialists. Although the senior graders agreed in most cases, in the atypical case of disagreement, the ground truth was obtained by a majority vote of the senior graders' quorum. The positive-label frequencies in this dataset were 32.68%, 51.71%, 42.93% and 12.68% for DV, IHRF, SDD and hDC, respectively.

The EchoNet-Dynamic Dataset

The EchoNet-Dynamic Dataset[50] was downloaded on 7 September 2022. The dataset contained 10,030 echocardiograms obtained from 10,030 different individuals who underwent echocardiography between 2006 and 2018. Each echocardiogram was labelled with a continuous number (between 0 and 1) representing an ejection fraction. The ejection fraction was obtained by a registered sonographer and further verified by a level 3 echocardiographer. The minimal ejection fraction in the dataset was 0.069, while the maximal ejection fraction was 0.97. The average ejection fraction was 0.558 with a standard deviation of 0.124. The dataset already set a random split for train, validation and test sets of sizes 7,465 (74.43%), 1,288 (12.84%) and 1,277 (12.73%), respectively. In contrast to the other datasets used in this study, the number of frames per video in the dataset was not constant but rather varied from 28 to 1,002 (with nearly 177 frames on average and a standard deviation of 58 frames). To standardize the data, we followed the same approach that the EchoNet paper authors took and sampled 32 equally spaced frames per volume[29].

The United Kingdom Biobank Dataset

The UKBB Dataset of MRI imaging with PDFF measurements was downloaded on 7 June 2022 from the UKBB repository[23]. The UKBB is a widely studied population-scale repository of phenotypic and genetic information for roughly half a million individuals. At the time of the study, the UKBB made available 16,876 PDFF measurements acquired from a subset of the 54,606 total hepatic-imaging MRIs. The MRI data of each individual consisted of an unordered series of 36 imaging scans in DICOM format at 284 by 288 resolution (in-plane pixel spacing 9.3 mm) acquired from a single breath-hold session. Of the data available, we identified a subset of 9,954 White British individuals who were unrelated and possessed both the hepatic MRI and PDFF measurement. The individuals were further divided into train, validation and test sets of sizes 5,972 (60%), 1,991 (20%) and 1,991 (20%), respectively.

The NoduleMNIST3D Dataset

The NoduleMNIST3D Dataset[62] is based on the Lung Image Database Consortium and Image Database Resource Initiative Dataset of volumetric-CT imaging[96]. The dataset contained 1,633 scans each (binary) labelled for nodule existence, with a positive-label frequency of 24.56%. The dataset was downloaded on 8 December 2023. The dataset has a predefined random split for train, validation and test sets of sizes 1,158 (70.91%), 165 (10.1%) and 310 (18.98%), with positive-label frequencies of 25.47%, 25.45% and 20.65%, respectively.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The 2D OCT dataset was downloaded from https://data.mendeley.com/datasets/rscbjbr9sj/3. The 3D OCT datasets are not publicly available owing to institutional data-use policy and to concerns about patient privacy. However, they are available from the authors upon reasonable request and with permission of the IRB. The echocardiogram dataset was downloaded from https://stanfordaimi.azurewebsites.net/datasets/834e1cd1-92f7-4268-9daa-d359198b310a. The MRI dataset was downloaded from https://www.ukbiobank.ac.uk under application number 33127. The 3D CT, the 2D CT and the 2D X-ray datasets were downloaded from https://medmnist.com.

Code availability

The code of SLIViT is available via the project's GitHub repository at https://github.com/cozygene/SLIViT.

References

1. Chiang, J. N. et al. Automated identification of incomplete and complete retinal epithelial pigment and outer retinal atrophy using machine learning. *Ophthalmol. Retina* **7**, 118–126 (2023).
2. Wong, T. Y., Liew, G. & Mitchell, P. Clinical update: new treatments for age-related macular degeneration. *Lancet* **370**, 204–206 (2007).
3. Gandhi, S. K. et al. The pathogenesis of acute pulmonary edema associated with hypertension. *N. Engl. J. Med.* **344**, 17–22 (2001).
4. Bloom, M. W. et al. Heart failure with reduced ejection fraction. *Nat. Rev. Dis. Primers* **3**, 17058 (2017).
5. Guindalini, R. S. C. et al. Intensive surveillance with biannual dynamic contrast-enhanced magnetic resonance imaging downstages breast cancer in BRCA1 mutation carriers. *Clin. Cancer Res.* **25**, 1786–1794 (2019).
6. Mann, R. M., Kuhl, C. K. & Moy, L. Contrast-enhanced MRI for breast cancer screening. *J. Magn. Reson. Imaging* **50**, 377–390 (2019).
7. Azizi, S. et al. Robust and data-efficient generalization of self-supervised machine learning for diagnostic imaging. *Nat. Biomed. Eng.* **7**, 756–779 (2023).
8. O'Shea, K. & Nash, R. An introduction to convolutional neural networks. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1511.08458 (2015).
9. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. *Commun. ACM* **60**, 84–90 (2017).
10. Esteva, A. et al. A guide to deep learning in healthcare. *Nat. Med.* **25**, 24–29 (2019).
11. Zhuang, F. et al. A comprehensive survey on transfer learning. In Proc. IEEE (ed. Setti, G.) 43–76 (IEEE, 2021).
12. McKinney, S. M. et al. International evaluation of an AI system for breast cancer screening. *Nature* **577**, 89–94 (2020).
13. Hannun, A. Y. et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nat. Med.* **25**, 65–69 (2019).
14. Rajpurkar, P. et al. Deep learning for chest radiograph diagnosis: a retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLoS Med.* **15**, e1002686 (2018).
15. Gulshan, V. et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* **316**, 2402–2410 (2016).
16. Deng, J. et al. ImageNet: a large-scale hierarchical image database. In *Proc. 2009 IEEE Conference on Computer Vision and Pattern Recognition* 248–255 (IEEE, 2009).
17. Tiu, E. et al. Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning. *Nat. Biomed. Eng.* **6**, 1399–1406 (2022).
18. Zhang, Y., Jiang, H., Miura, Y., Manning, C. D. & Langlotz, C. P. Contrastive learning of medical visual representations from paired images and text. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2010.00747 (2020).
19. Xie, Y., Zhang, J., Xia, Y. & Wu, Q. UniMiSS: Universal Medical Self-Supervised learning via breaking dimensionality barrier. In *Proc. European Conference on Computer Vision* (eds. Avidan, S. et al.) 558–575 (Springer, 2022).
20. Azizi, S. et al. Big self-supervised models advance medical image classification. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2101.05224 (2021).
21. Wu, Z. et al. OCT signs of early atrophy in age-related macular degeneration: interreader agreement: classification of atrophy meetings report 6. *Ophthalmol. Retina* **6**, 4–14 (2022).
22. Huang, S.-C. et al. Self-supervised learning for medical image classification: a systematic review and implementation guidelines. *npj Digit. Med.* **6**, 74 (2023).
23. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
24. Willemink, M. J. et al. Preparing medical imaging data for machine learning. *Radiology* **295**, 4–15 (2020).
25. Tran, D. et al. A closer look at spatiotemporal convolutions for action recognition. In *Proc. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* 6450–6459 (IEEE, 2018).
26. Arnab, A. et al. ViViT: a video vision transformer. In *Proc. 2021 IEEE/CVF International Conference on Computer Vision (ICVV)* 6816–6826 (IEEE, 2021).
27. Zhu, H., Chen, B. & Yang, C. Understanding why ViT trains badly on small datasets: an intuitive perspective. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2302.03751 (2023).
28. Rakocz, N. et al. Automated identification of clinical features from sparsely annotated 3-dimensional medical imaging. *npj Digit. Med.* **4**, 44 (2021).
29. Ghorbani, A. et al. Deep learning interpretation of echocardiograms. *npj Digit. Med.* **3**, 10 (2020).
30. Gupta, U. et al. Transferring models trained on natural images to 3D MRI via position encoded slice models. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2303.01491 (2023).
31. Witowski, J. et al. Improving breast cancer diagnostics with deep learning for MRI. *Sci. Transl. Med.* **14**, eabo4802 (2022).
32. Yang, M., Huang, X., Huang, L. & Cai, G. Diagnosis of Parkinson's disease based on 3D ResNet: the frontal lobe is crucial. *Biomed. Signal Process. Control* **85**, 104904 (2023).
33. Zou, Q. et al. Three-dimensional ultrasound image reconstruction based on 3D-ResNet in the musculoskeletal system using a 1D probe: ex vivo and in vivo feasibility studies. *Phys. Med. Biol.* **68**, 165003 (2023).
34. Turnbull, R. Using a 3D ResNet for detecting the presence and severity of COVID-19 from CT scans. In *Proc. Computer Vision – ECCV 2022 Workshops* (eds Karlinsky, L. et al.) 663–676 (Springer, 2023).
35. Caron, M. et al. Emerging properties in self-supervised vision transformers. In *Proc. 2021 IEEE/CVF International Conference on Computer Vision (ICVV)* 9630–9640 (IEEE, 2021).
36. Zhou, H.-Y., Lu, C., Yang, S., Han, X. & Yu, Y. Preservational learning improves self-supervised medical image models by reconstructing diverse contexts. In *Proc. 2021 IEEE/CVF International Conference on Computer Vision (ICVV)* 3479–3489 (IEEE, 2021).
37. Xie, Y., Zhang, J., Liao, Z., Xia, Y. & Shen, C. PGL: prior-guided local self-supervised learning for 3D medical image segmentation. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2011.12640 (2020).

38. Chen, X., Fan, H., Girshick, R. & He, K. Improved baselines with momentum contrastive learning. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2003.04297 (2020).
39. Chen, X., Xie, S. & He, K. An empirical study of training self-supervised vision transformers. In *Proc. 2021 IEEE/CVF International Conference on Computer Vision (ICVV)* 9620–9629 (IEEE, 2021).
40. Liu, Z. et al. A ConvNet for the 2020s. In *Proc. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 11966–11976 (IEEE, 2022).
41. Dosovitskiy, A. et al. An image is worth 16x16 words: transformers for image recognition at scale. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2010.11929 (2021)
42. Gupta, U., Lam, P. K., Ver Steeg, G. & Thompson, P. M. Improved brain age estimation with slice-based set networks. In *Proc. 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)* 840–844 (IEEE, 2021).
43. Zeng, Y. et al. A 2.5D deep learning-based method for drowning diagnosis using post-mortem computed tomography. *IEEE J. Biomed. Health Inform.* **27**, 1026–1035 (2023).
44. Schlemper, J. et al. Attention gated networks: learning to leverage salient regions in medical images. *Med. Image Anal.* **53**, 197–207 (2019).
45. Bertasius, G., Wang, H. & Torresani, L. Is space-time attention all you need for video understanding? In *Proc. 38th International Conference on Machine Learning (ICML)* (2021).
46. Neimark, D., Bar, O., Zohar, M. & Asselmann, D. Video transformer network. In *Proc. IEEE/CVF International Conference on Computer Vision (ICVV)* 3156–3165 (2021).
47. Kermany, D. S. et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell* **172**, 1122–1131.e9 (2018).
48. Wong, W. L. et al. Global prevalence of age-related macular degeneration and disease burden projection for 2020 and 2040: a systematic review and meta-analysis. *Lancet Glob. Health* **2**, e106–e116 (2014).
49. Hirabayashi, K. et al. OCT risk factors for development of atrophy in eyes with intermediate age-related macular degeneration. *Ophthalmol. Retina* **7**, 253–260 (2023).
50. Ouyang, D. et al. EchoNet-Dynamic: a large new cardiac motion video data resource for medical machine learning. In *Proc. 33rd International Conference on Neural Information Processing Systems* (eds Wallach, H. M.) (Curran Associates Inc., 2019).
51. Ziaeian, B. & Fonarow, G. C. Epidemiology and aetiology of heart failure. *Nat. Rev. Cardiol.* **13**, 368–378 (2016).
52. Klapholz, M. et al. Hospitalization for heart failure in the presence of a normal left ventricular ejection fraction: results of the New York Heart Failure Registry. *J. Am. Coll. Cardiol.* **43**, 1432–1438 (2004).
53. Dunlay, S. M., Roger, V. L. & Redfield, M. M. Epidemiology of heart failure with preserved ejection fraction. *Nat. Rev. Cardiol.* **14**, 591–602 (2017).
54. Idilman, I. S. et al. Hepatic steatosis: quantification by proton density fat fraction with MR imaging versus liver biopsy. *Radiology* **267**, 767–775 (2013).
55. Jung, J. et al. Direct comparison of quantitative US versus controlled attenuation parameter for liver fat assessment using MRI proton density fat fraction as the reference standard in patients suspected of having NAFLD. *Radiology* **304**, 75–82 (2022).
56. Runge, J. H. et al. MR spectroscopy-derived proton density fat fraction is superior to controlled attenuation parameter for detecting and grading hepatic steatosis. *Radiology* **286**, 547–556 (2018).
57. Schawkat, K. et al. Preoperative evaluation of pancreatic fibrosis and lipomatosis: correlation of magnetic resonance findings with histology using magnetization transfer imaging and multigradient echo magnetic resonance imaging. *Invest. Radiol.* **53**, 720–727 (2018).
58. Kühn, J.-P. et al. Pancreatic steatosis demonstrated at MR imaging in the general population: clinical relevance. *Radiology* **276**, 129–136 (2015).
59. Patel, N. S. et al. Insulin resistance increases MRI-estimated pancreatic fat in nonalcoholic fatty liver disease and normal controls. *Gastroenterol. Res. Pract.* **2013**, 498296 (2013).
60. Trout, A. T. et al. Relationship between abdominal fat stores and liver fat, pancreatic fat, and metabolic comorbidities in a pediatric population with non-alcoholic fatty liver disease. *Abdom. Radiol.* **44**, 3107–3114 (2019).
61. Covarrubias, Y. et al. Pilot study on longitudinal change in pancreatic proton density fat fraction during a weight-loss surgery program in adults with obesity. *J. Magn. Reson. Imaging* **50**, 1092–1102 (2019).
62. Yang, J. et al. MedMNIST v2 – a large-scale lightweight benchmark for 2D and 3D biomedical image classification. *Sci. Data* **10**, 41 (2023).
63. Halder, A., Dey, D. & Sadhu, A. K. Lung nodule detection from feature engineering to deep learning in thoracic CT images: a comprehensive review. *J. Digit. Imaging* **33**, 655–677 (2020).
64. Mazzone, P. J. & Lam, L. Evaluating the patient with a pulmonary nodule: a review. *JAMA* **327**, 264–273 (2022).
65. Esteva, A. et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115–118 (2017).
66. Huang, Z., Bianchi, F., Yuksekgonul, M., Montine, T. J. & Zou, J. A visual-language foundation model for pathology image analysis using medical Twitter. *Nat. Med.* https://doi.org/10.1038/s41591-023-02504-3 (2023).
67. Liu, Y. et al. A deep learning system for differential diagnosis of skin diseases. *Nat. Med.* **26**, 900–908 (2020).
68. Guan, H., Wang, L., Yao, D., Bozoki, A. & Liu, M. Learning transferable 3D-CNN for MRI-based brain disorder classification from scratch: an empirical study. In *Proc. Machine Learning in Medical Imaging* (eds. Lian, C. et al.) 10–19 (Springer, 2021).
69. Mustafa, B. et al. Supervised transfer learning at scale for medical imaging. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2101.05913 (2021).
70. Raghu, M., Zhang, C., Kleinberg, J. & Bengio, S. Transfusion: understanding transfer learning for medical imaging. In *Proc. 33rd International Conference on Neural Information Processing Systems* (eds Wallach, H. M.) (Curran Associates Inc., 2019).
71. Zhou, Y. et al. A foundation model for generalizable disease detection from retinal images. *Nature* **622**, 156–163 (2023).
72. Newell, A. & Deng, J. How useful is self-supervised pretraining for visual tasks? In *Proc. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 7343–7352 (IEEE, 2020).
73. Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. A simple framework for contrastive learning of visual representations. In *Proc. 37th International Conference on Machine Learning* (eds Daumé, H. & Singh, A.) (JMLR, 2020).
74. He, K., Fan, H., Wu, Y., Xie, S. & Girshick, R. Momentum contrast for unsupervised visual representation learning. In *Proc. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition* 9726–9735 (IEEE, 2020).
75. Mitrovic, J., McWilliams, B., Walker, J., Buesing, L. & Blundell, C. Representation learning via invariant causal mechanisms. In *Proc. International Conference on Learning Representations* (2020).
76. Zbontar, J., Jing, L., Misra, I., LeCun, Y. & Deny, S. Barlow twins: self-supervised learning via redundancy reduction. In *Proc. International Conference on Machine Learning* (2021).

77. Kornblith, S., Norouzi, M., Lee, H. & Hinton, G. Similarity of neural network representations revisited. In *Proc. 36th International Conference on Machine Learning* (eds Chaudhuri, K. & Salakhutdinov, R.) (Curran Associates, Inc., 2019).
78. Taleb, A. et al. 3D self-supervised methods for medical imaging. In *Proc. 34th International Conference on Neural Information Processing Systems* (Curran Associates, Inc., 2020).
79. Tang, Y. et al. Self-supervised pre-training of Swin transformers for 3D medical image analysis. In *Proc. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition* 20698–20708 (IEEE, 2022).
80. He, K. et al. Masked autoencoders are scalable vision learners. In *Proc. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition* 15979–15988 (IEEE, 2022).
81. Woo, S. et al. ConvNeXt V2: co-designing and scaling ConvNets with masked autoencoders. In *Proc. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition* 16133–16142 (IEEE, 2023).
82. Kadambi, A. Achieving fairness in medical devices. *Science* **372**, 30–31 (2021).
83. Chen, R. J. et al. Algorithmic fairness in artificial intelligence for medicine and healthcare. *Nat. Biomed. Eng.* **7**, 719–742 (2023).
84. Paszke, A. et al. PyTorch: an imperative style, high-performance deep learning library. In *Proc. 33rd International Conference on Neural Information Processing Systems* (eds Wallach, H. M. et al.) (Curran Associates Inc., 2019).
85. Howard, J. & Gugger, S. fastai: a layered API for deep learning. *Information*, **11**, 108 (2020).
86. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
87. Hendrycks, D. & Gimpel, K. Gaussian error linear units (GELUs). Preprint at *arXiv* https://doi.org/10.48550/arXiv.1606.08415 (2016).
88. Smith, L. N. Cyclical learning rates for training neural networks. In *Proc. 2017 IEEE Winter Conference on Applications of Computer Vision (WACV)* 464–472 (IEEE, 2017).
89. Smith, L. N. & Topin, N. Super-convergence: very fast training of neural networks using large learning rates. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1708.07120 (2018).
90. Rajkomar, A. et al. Scalable and accurate deep learning with electronic health records. *npj Digit. Med.* **1**, 18 (2018).
91. Kay, W. et al. The kinetics human action video dataset. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1705.06950 (2017).
92. Ferris, F. L. et al. Clinical classification of age-related macular degeneration. *Ophthalmology* **120**, 844–851 (2013).
93. Nassisi, M. et al. OCT risk factors for development of late age-related macular degeneration in the fellow eyes of patients enrolled in the HARBOR study. *Ophthalmology* **126**, 1667–1674 (2019).
94. Lei, J., Balasubramanian, S., Abdelfattah, N. S., Nittala, M. G. & Sadda, S. R. Proposal of a simple optical coherence tomography-based scoring system for progression of age-related macular degeneration. *Graefes Arch. Clin. Exp. Ophthalmol.* **255**, 1551–1558 (2017).
95. Nittala, M. G. et al. Amis Eye Study: baseline spectral domain optical coherence tomography characteristics of age-related macular degeneration. *Retina* **39**, 1540–1550 (2019).
96. Armato, S. G. et al. The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): a completed reference database of lung nodules on CT scans. *Med. Phys.* **38**, 915–931 (2011).

Acknowledgements

This work was supported by a National Institutes of Health (NIH)/National Institute of General Medical Sciences grant 5R25GM135043, an NIH/National Institute of Biomedical Imaging and Bioengineering grant R01EB035028, NIH/National Eye Institute grants R01EY023164 and 1R01EY030614, and an Unrestricted Grant from Research to Prevent Blindness. This research was conducted using the UK Biobank Resource under application number 33127. We also acknowledge the participants of the UCLA Computational Genomics Summer Institute (supported by the aforementioned NIH/National Institute of General Medical Sciences grant) for stimulating discussions.

Author contributions

O.A., B.D., N.R., G.C., U.A., M.G.N., B.Z., S.S., J.N.C., S.R.S. and E.H. contributed to the conception and design of the work. O.A., B.D., G.C., U.A., M.G.N., A.R., Z.J.C., Y.W., K.H., S.V., L.T., F.C., A.V., A.K., S.L., D.O., L.A., V.H., S.F.-A., H.E., C.C.W., S.R.S. and E.H. contributed to data acquisition. O.A., B.D., N.R., G.C., B.Z., N.Z., I.G., J.N.C., S.R.S. and E.H. contributed to the evaluation of the work. O.A., B.D., N.R., G.C., U.A., M.G.N., P.T., I.G., S.S., J.N.C., S.R.S. and E.H. contributed to the analysis and interpretation of the data. O.A., B.D., N.R., G.C., U.A., A.R., M.C., E.R., C.W.A., N.Z., I.G., S.S., J.N.C., S.R.S. and E.H. contributed to drafting and revising the paper. S.R.S. and E.H. contributed equally as co-advisers. All authors read and approved the final version of the paper.

Competing interests

E.H. has an affiliation with Optum. S.R.S. has affiliations with Abbvie/Allergan, Alexion, Amgen, Apellis, ARVO, Astellas, Bayer, Biogen, Boerhinger Ingelheim, Carl Zeiss Meditec, Centervue, Character, Eyepoint, Heidelberg, iCare, IvericBio, Jannsen, Macula Society, Nanoscope, Nidek, NotalVision, Novartis, Optos, OTx, Pfizer, Regeneron, Roche, Samsung Bioepis and Topcon. The other authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41551-024-01257-9.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41551-024-01257-9.

Correspondence and requests for materials should be addressed to Oren Avram, Srinivas R. Sadda or Eran Halperin.

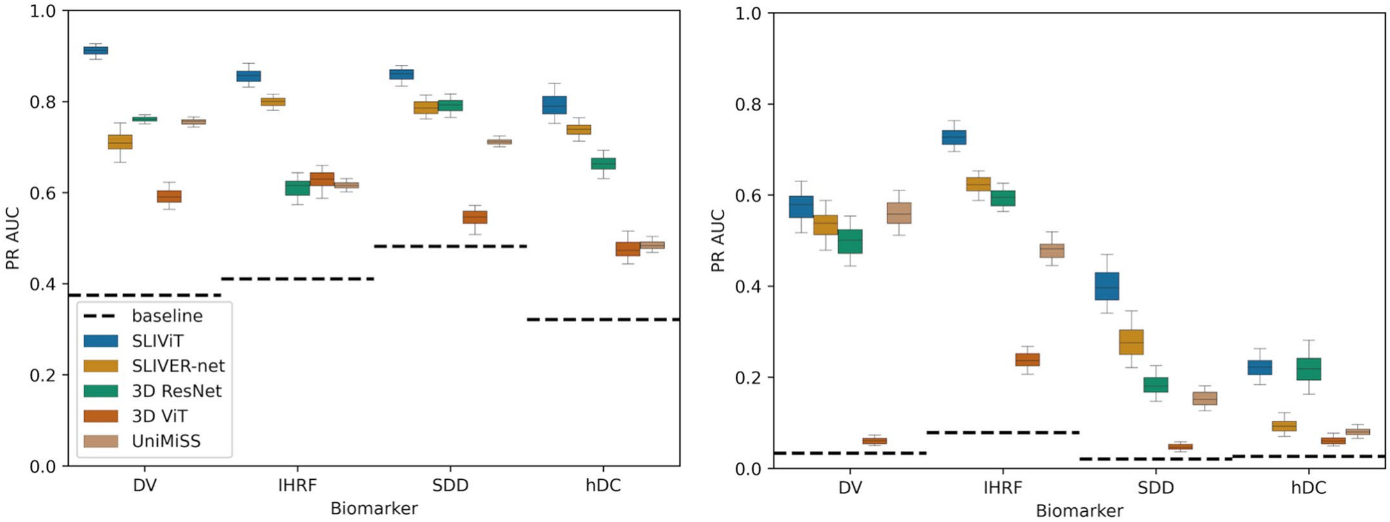
Peer review information *Nature Biomedical Engineering* thanks Tianyu Zhang, Yukun Zhou and the two other, anonymous, reviewers for their contribution to the peer review of this work.

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

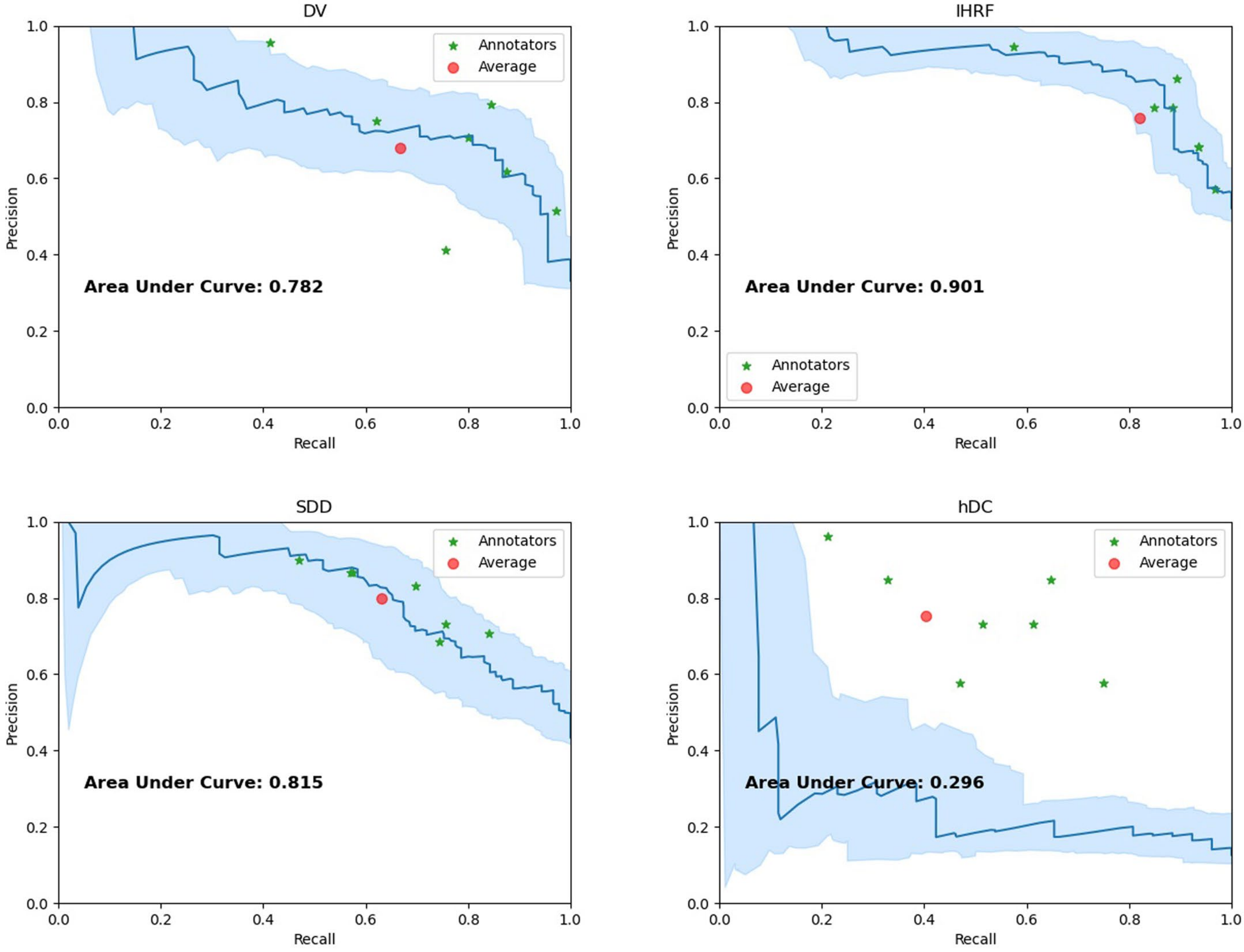
Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Oren Avram[1,2,3,17] ✉, Berkin Durmus[2,17], Nadav Rakocz[2], Giulia Corradetti[4,5], Ulzee An[1,2], Muneeswar G. Nittala[4,5], Prerit Terway[1,2], Akos Rudas[1], Zeyuan Johnson Chen[2], Yu Wakatsuki[4], Kazutaka Hirabayashi[4], Swetha Velaga[4], Liran Tiosano[4,6], Federico Corvi[4], Aditya Verma[4,7], Ayesha Karamat[4], Sophiana Lindenberg[4], Deniz Oncel[4], Louay Almidani[4], Victoria Hull[4], Sohaib Fasih-Ahmad[4], Houri Esmaeilkhanian[4], Maxime Cannesson[3], Charles C. Wykoff[8,9], Elior Rahmani[1], Corey W. Arnold[10,11,12], Bolei Zhou[1,2], Noah Zaitlen[13,14], Ilan Gronau[15], Sriram Sankararaman[1,2,14], Jeffrey N. Chiang[1,16], Srinivas R. Sadda[4,5] ✉ & Eran Halperin[2] ✉

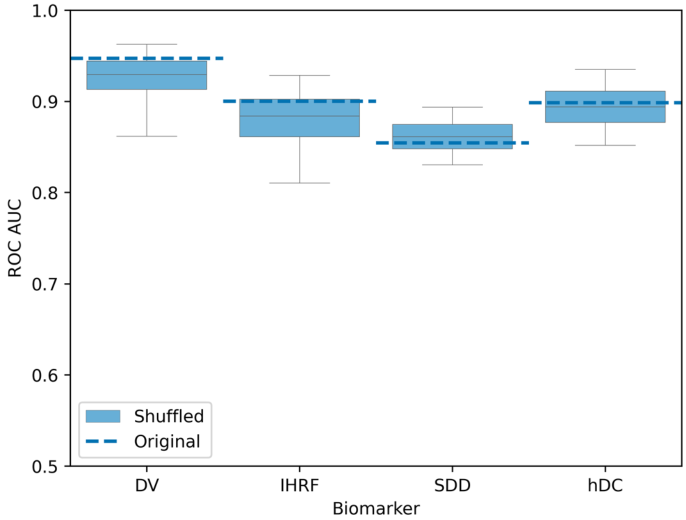
[1]Department of Computational Medicine, University of California, Los Angeles, Los Angeles, CA, USA. [2]Department of Computer Science, University of California, Los Angeles, Los Angeles, CA, USA. [3]Department of Anesthesiology and Perioperative Medicine, University of California, Los Angeles, Los Angeles, CA, USA. [4]Doheny Eye Institute, University of California, Los Angeles, Pasadena, CA, USA. [5]Department of Ophthalmology, University of California, Los Angeles, Los Angeles, CA, USA. [6]Department of Ophthalmology, Hadassah-Hebrew University Medical Center, Jerusalem, Israel. [7]Department of Ophthalmology and Visual Sciences, University of Louisville, Louisville, KY, USA. [8]Retina Consultants of Texas, Retina Consultants of America, Houston, TX, USA. [9]Blanton Eye Institute, Houston Methodist Hospital, Houston, TX, USA. [10]Department of Radiology, University of California, Los Angeles, Los Angeles, CA, USA. [11]Department of Bioengineering, University of California, Los Angeles, Los Angeles, CA, USA. [12]Department of Pathology, University of California, Los Angeles, Los Angeles, CA, USA. [13]Department of Neurology, University of California, Los Angeles, Los Angeles, CA, USA. [14]Department of Human Genetics, University of California, Los Angeles, Los Angeles, CA, USA. [15]School of Computer Science, Reichman University, Herzliya, Israel. [16]Department of Neurosurgery, University of California, Los Angeles, Los Angeles, CA, USA. [17]These authors contributed equally: Oren Avram, Berkin Durmus. ✉e-mail: orenavram@gmail.com; ssadda@doheny.org; ehalperin@cs.ucla.edu



Extended Data Fig. 1 | PR AUC comparison of five models in four single-task AMD-biomarker classification problems when trained on less than 700 OCT volumes. Shown are the PR AUC as an alternative scoring metric for the OCT experiments shown in Fig. 3. The left panel shows the performance when trained and tested on the Houston Dataset (see Supplementary Table 1). The right panel shows the performance when trained on the Houston Dataset and tested on the SLIVER-net Dataset (see Supplementary Table 2). The dashed lines represent the corresponding biomarker's positive-label prevalence, which is the expected performance of a random model. Box plot whiskers represent a 90% CI.

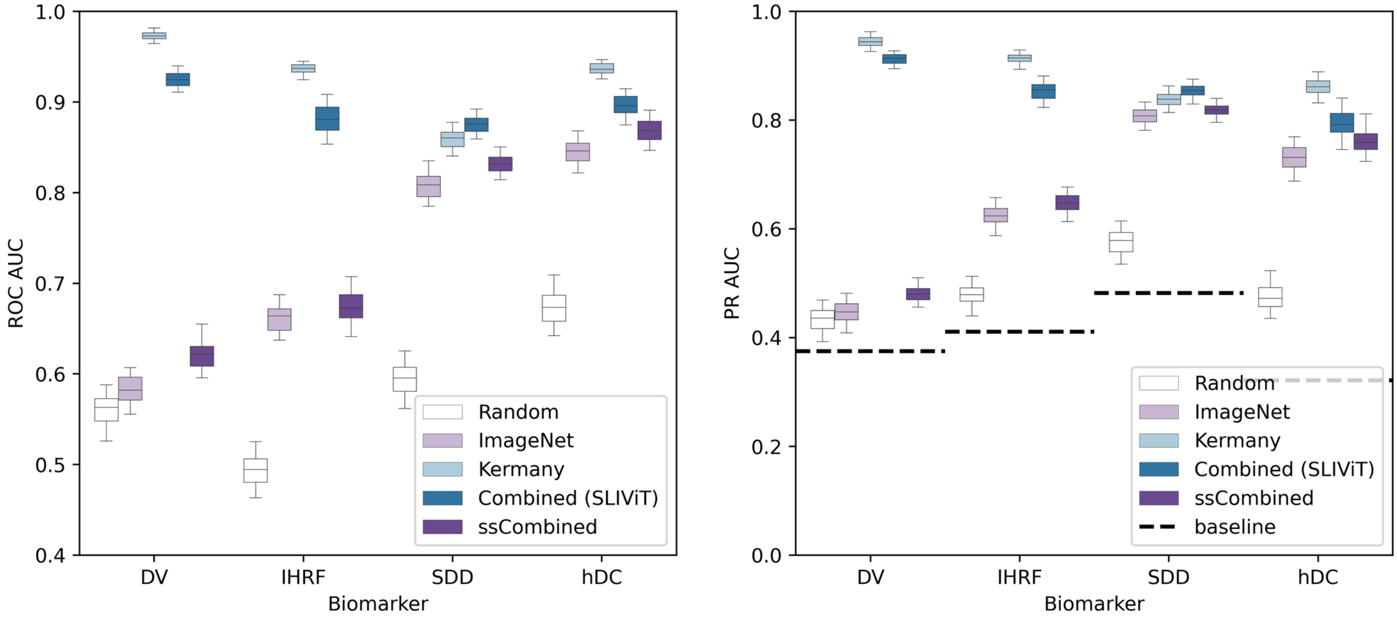


Extended Data Fig. 2 | Precision-recall performance compared to clinical retina specialists' assessment. Shown are the PR curves (blue) of SLIViT as an alternative scoring metric for the OCT experiments shown in Fig. 5. SLIViT was trained using less than 700 OCT volumes (Houston Dataset) and tested on an independent dataset (Pasadena Dataset). In each panel, the light-blue shaded area represents a 90% CI for SLIViT's performance, the red dot represents the retina clinical specialists' average performance, and the green asterisks correspond to the retina clinical specialists' assessments. Two of the clinical specialists obtained the exact same performance score for IHRF classification.

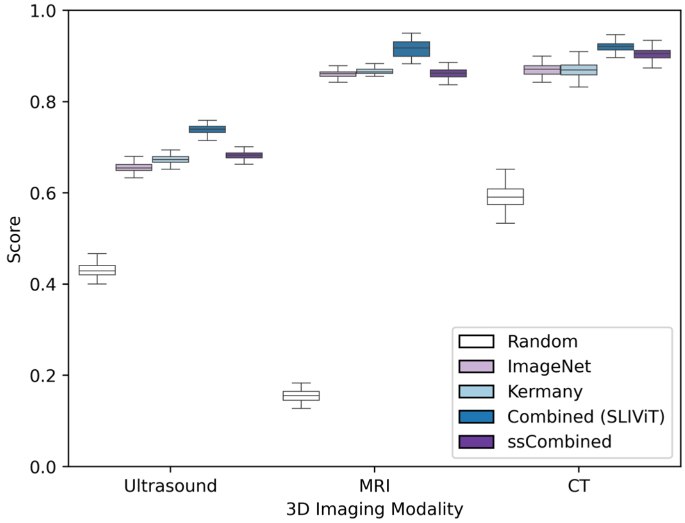


Extended Data Fig. 3 | SLIViT's performance in a frame-shuffling experiment. Shown are the ROC AUC scores distribution of 101 SLIViT models in four single-task classification problems of AMD high-risk factors (DV, IHRF, SDD, and hDC) trained on volumetric-OCT dataset. One model was trained on the OCT dataset in its original form, while the other 100 models were trained on randomly shuffled copies of the dataset. The performance ranks of the former model (Original) compared to the performance distribution of the latter models (Shuffled) were 22, 34, 56, and 47 for DV, IHRF, SDD, and hDC, respectively. The expected performance of a random classifier is 0.5. Box plot whiskers extend to the 5th and the 95th ranked models (out of the 100 shuffled models' performance distribution).

https://doi.org/10.1038/s41551-024-01257-9

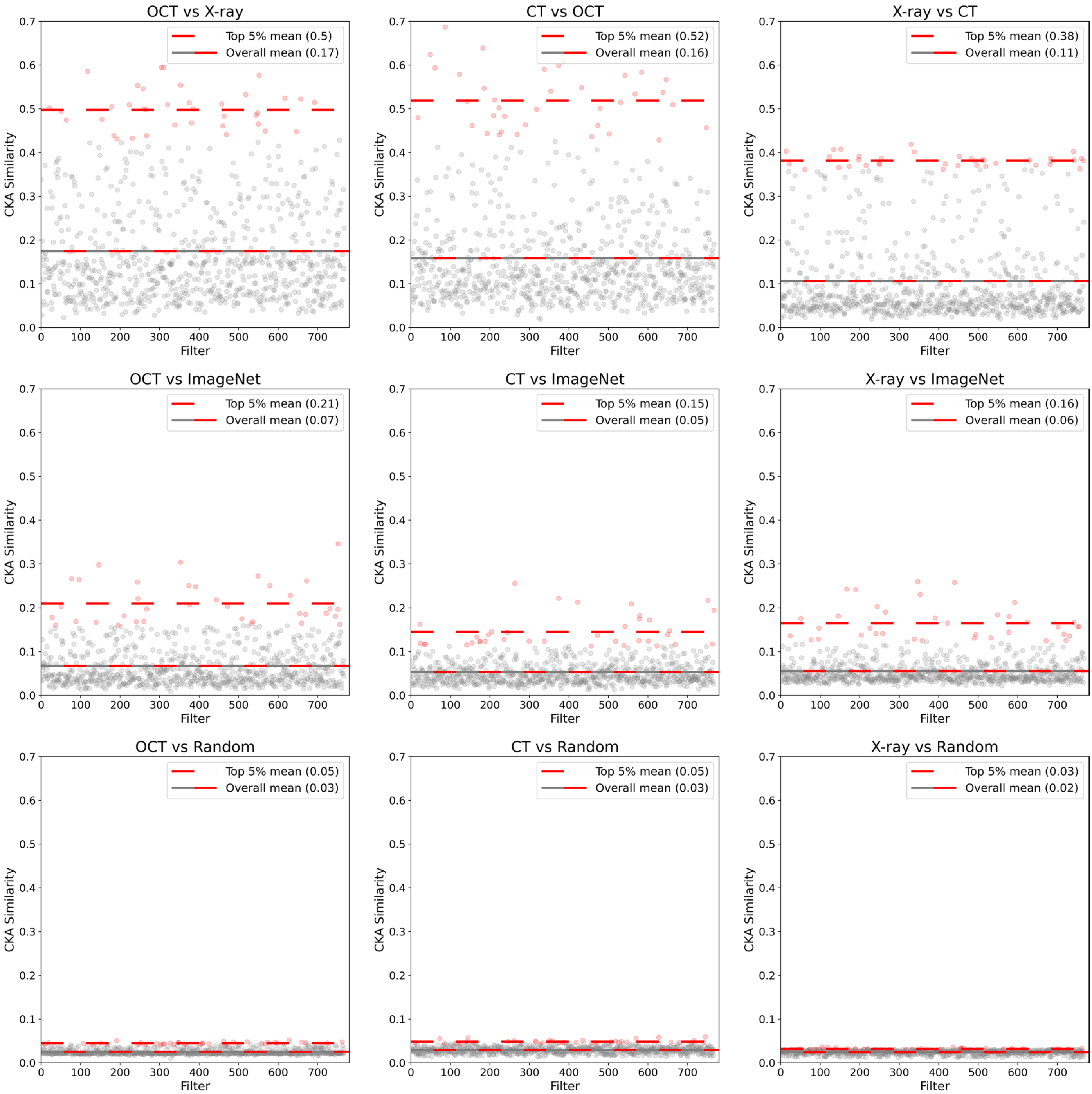


Extended Data Fig. 4 | ImageNet and OCT B-scans pre-training contribution for OCT-related downstream learning tasks. Shown are the ROC (left) and PR (right) AUC scores across different fine-tuned models for volumetric-OCT classification tasks initialized with five different sets of weights. Combined, the proposed SLIViT's initialization, is ImageNet weights initialization followed by supervised pre-training on the Kermany Dataset. ssCombined is an ImageNet weights initialization followed by self-supervised pre-training on an unlabeled version of the Kermany Dataset. The expected ROC AUC score of a random model is 0.5. The dashed lines represent the corresponding biomarker's positive-label prevalence, which is the expected PR AUC score of a random model. Box plot whiskers represent a 90% CI.

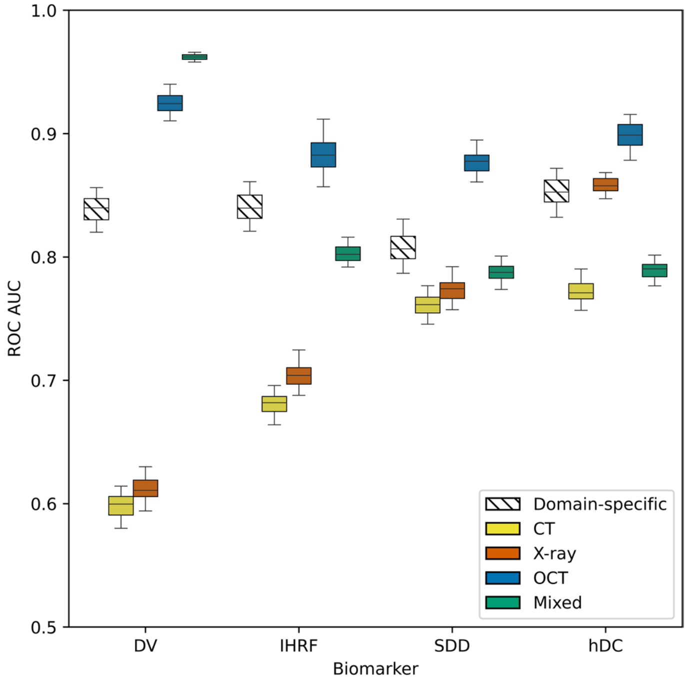


Extended Data Fig. 5 | ImageNet and OCT B-scans pre-training contribution for non-OCT-related downstream learning tasks. Shown are the performance scores for the volumetric ultrasound and MRI regression tasks (R^2) and the volumetric CT classification task (ROC AUC) initialized with five different sets of weights. Combined, the proposed SLIViT's initialization, is ImageNet weights initialization followed by supervised pre-training on the Kermany Dataset. ssCombined is an ImageNet weights initialization followed by self-supervised pre-training on an unlabeled version of the Kermany Dataset. The expected R^2 and ROC AUC of a random model are 0 and 0.5, respectively. Box plot whiskers represent a 90% CI.

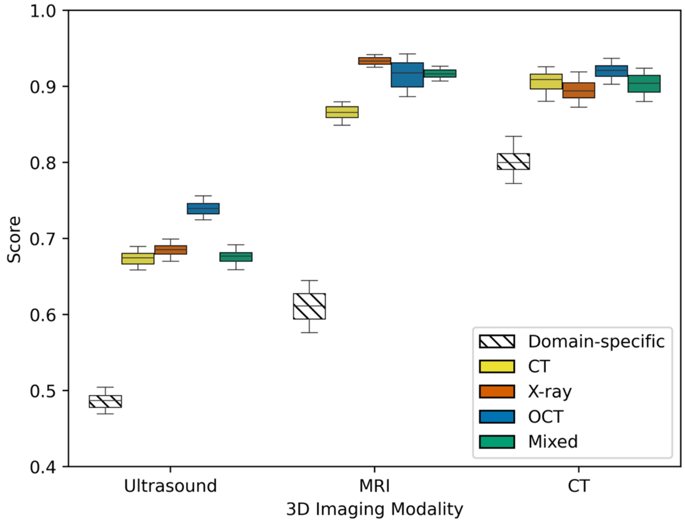
https://doi.org/10.1038/s41551-024-01257-9



Extended Data Fig. 6 | Feature similarity analysis between various pre-trained backbone projections. Shown are nine scatterplots of similarity analysis (CKA) when comparing the projections of a biomedical-imaging dataset induced by different biomedical-imaging pre-trained backbones. Each panel corresponds to a different pair of pre-trained backbones (upper- biomedical pairs; middle- biomedical and ImageNet pairs; lower- biomedical and random pairs). In each panel, each of the 768 dots represents the similarity score computed for the projections induced by the corresponding filter. A dot is red if it falls within the top 5% scores (and gray otherwise). The dashed lines show the average score measured for the color-corresponding set of dots.



Extended Data Fig. 7 | 2D biomedical-imaging pre-training performance contribution for 3D OCT-related downstream learning tasks. Shown are the ROC AUC scores on four volumetric-OCT single-task classification problems. Four SLIViT models were evaluated in every classification problem. Each SLIViT model was initialized with ImageNet weights and then pre-trained on a 2D biomedical-imaging dataset of a different modality. The considered modalities were CT, X-ray, OCT, and Mixed (containing all the images from the CT, X-ray, and OCT datasets). SLIVER-net's performance (Domain-specific) is borrowed from Fig. 3. The expected performance of a random model is 0.5. Box plot whiskers represent a 90% CI.



Extended Data Fig. 8 | 2D biomedical-imaging pre-training performance contribution for 3D non-OCT-related downstream learning tasks. Shown are the performance scores for the volumetric ultrasound and MRI regression tasks (R^2) and the volumetric CT classification task (ROC AUC). Four SLIViT models were evaluated in every learning problem. Each SLIViT model was initialized with ImageNet weights and then pre-trained on a 2D biomedical-imaging dataset of a different modality. The considered modalities were CT, X-ray, OCT, and Mixed (containing all the images from the CT, X-ray, and OCT datasets). The performance scores of the domain-specific methods were borrowed from Fig. 2. The expected R^2 and ROC AUC of a random model are 0 and 0.5, respectively. Box plot whiskers represent a 90% CI.

nature portfolio

Corresponding author(s):	Oren Avram, Srinivas R. Sadda and Eran Halperin
Last updated by author(s):	Jun 3, 2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
□	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
□	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
□	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	□	A description of all covariates tested
☒	□	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
□	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
□	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	□	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	□	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
□	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used to collect the data.
Data analysis	The deep-learning model was implemented in Python 3.8 using the libraries listed on the project's repository. The code of SLIViT is available in the project’s GitHub repository at https://github.com/cozygene/SLIViT.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The 2D OCT dataset was downloaded from https://data.mendeley.com/datasets/rscbjbr9sj/3. The 3D OCT datasets are not publicly available owing to institutional data-use policy and to concerns about patient privacy. However, they are available from the authors upon reasonable request and with permission of the IRB. The echocardiogram dataset was downloaded from https://stanfordaimi.azurewebsites.net/datasets/834e1cd1-92f7-4268-9daa-d359198b310a. The MRI dataset was downloaded from https://www.ukbiobank.ac.uk under application number 33127. The 3D CT, the 2D CT, and the 2D X-ray datasets were downloaded from https://medmnist.com.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender

This information was not collected.

Reporting on race, ethnicity, or other socially relevant groupings

No socially constructed groupings were used in the analyses.

Population characteristics

See above.

Recruitment

A waiver of informed consent was granted, given the retrospective nature of the study. Patient imaging data were de-identified.

Ethics oversight

The OCT studies were reviewed and approved by the Ethics Committee of Retina Consultants Texas (Houston Methodist Hospital, Pro00020661:1 “Retrospective Prospective Analysis of Retinal Diseases”) and by the IRB of the University of California, Los Angeles (UCLA IRB # 15-000083).

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences　□ Behavioural & social sciences　□ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size

We used volumetric-imaging data from eight different sources. Sample sizes were ranging from 205 (Pasadena Dataset) to 10,030 (EchoNet-Dynamic Dataset).

Data exclusions

No data were excluded.

Replication

Multiple replications of the prediction experiments were performed to replicate the results from the model, and confidence intervals were used to assess the variation in the results (where applicable).

Randomization

All training, validation and test sets were randomly split according to common machine-learning best practices.

Blinding

Blinding was not relevant to the study, as it was a retrospective analysis.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study	
☒	□	Antibodies
☒	□	Eukaryotic cell lines
☒	□	Palaeontology and archaeology
☒	□	Animals and other organisms
☒	□	Clinical data
☒	□	Dual use research of concern
☒	□	Plants

Methods

n/a	Involved in the study	
☒	□	ChIP-seq
☒	□	Flow cytometry
☒	□	MRI-based neuroimaging