SLiViT: a general AI framework for clinical-feature diagnosis from limited 3D biomedical-imaging data

Oren Avram
University of California, Los Angeles  https://orcid.org/0000-0003-1984-2139

Berkin Durmus
University of California, Los Angeles

Nadav Rakocz
University of California, Los Angeles

Giulia Corradetti
Doheny Eye Institute

Ulzee An
University of California, Los Angeles

Muneeswar Nitalla
Doheny Eye Institute

Ákos Rudas
University of California Los Angeles  https://orcid.org/0000-0003-4346-8239

Yu Wakatsuki
Doheny Eye Institute

Kazutaka Hirabayashi
Doheny Eye Institute

Swetha Velaga
Doheny Eye Institute

Liran Tiosano
Doheny Eye Institute

Federico Corvi
Doheny Eye Institute  https://orcid.org/0000-0002-2661-5500

Aditya Verma
Doheny Eye Institute

Ayesha Karamat
Doheny Eye Institute

Sophiana Lindenberg
Doheny Eye Institute

Deniz Oncel
Doheny Eye Institute

Louay Almidani
Doheny Eye Institute

Victoria Hull
Doheny Eye Institute

Sohaib Fasih-Ahmad
Doheny Eye Institute

Houri Esmaeilkhanian
Doheny Eye Institute

Charles Wykoff
Retina Consultants of Texas

Elior Rahmani
University of California, Los Angeles

Corey Arnold
University of California, Los Angeles

Bolei Zhou
University of California, Los Angeles

Noah Zaitlen
University of California, Los Angeles

Ilan Gronau
Reichman University

Sriram Sankararaman
University of California, Los Angeles

Jeffrey Chiang
UCLA  https://orcid.org/0000-0002-6843-1355

Srinivas Sadda
Doheny Eye Institute

Eran Halperin (✉ eranhalperin@gmail.com)
University of California, Los Angeles

---

**Article**

**Keywords:**

**Posted Date:** June 21st, 2023

**DOI:** https://doi.org/10.21203/rs.3.rs-3044914/v1

**License:** ☕️ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)
Additional Declarations: Yes there is potential Competing Interest. Prof. Halperin has an affiliation with Optum
SLIViT: a general AI framework for clinical-feature diagnosis from limited 3D biomedical-imaging data

Oren Avram*,1,2,3, Berkin Durmus*,2, Nadav Rakocz2, Giulia Corradetti4,5, Ulzee An1,2, Muneeswar G. Nitalla4,5, Akos Rudas1, Yu Wakatsuki4, Kazutaka Hirabayashi4, Swetha Velaga4, Liran Tiosano4, Federico Corvi4, Aditya Verma4,6, Ayesha Karamat4, Sophiana Lindenberg4, Deniz Oncel4, Louay Almidani4, Victoria Hull4, Sohaib Fasih-Ahmad4, Houri Esmaeilkhanian4, Charles C. Wykoff7, Elior Rahmani1, Corey W. Arnold8,9,10, Bolei Zhou2, Noah Zaitlen11,12, Ilan Gronau13, Sriram Sankararaman1,2,12, Jeffrey N. Chiang1, Srinivas R. Sadda+,4,5, Eran Halperin*,1,2,3,12

1 Department of Computational Medicine, University of California Los Angeles, Los Angeles, California, United States of America
2 Department of Computer Science, University of California Los Angeles, Los Angeles, California, United States of America
3 Department of Anesthesiology and Perioperative Medicine, University of California Los Angeles, Los Angeles, California, United States of America
4 Doheny Eye Institute, Pasadena, California, United States of America
5 Department of Ophthalmology, University of California Los Angeles, Los Angeles, California, United States of America
6 Department of Ophthalmology and Visual Sciences, University of Louisville, Kentucky, United States of America
7 Retina Consultants of Texas, Retina Consultants of America, Houston, Texas
8 Departments of Radiology, University of California Los Angeles, Los Angeles, California, United States of America
9 Departments of Bioengineering, University of California Los Angeles, Los Angeles, California, United States of America
10 Departments of Pathology, University of California Los Angeles, Los Angeles, California, United States of America
11 Department of Neurology, University of California Los Angeles, Los Angeles, California, United States of America
12 Department of Human Genetics, University of California Los Angeles, Los Angeles, California, United States of America
13 School of Computer Science, Reichman University, Herzliya, Israel

*Equal contribution
+Joint supervision
Abstract

We present SLIViT, a deep-learning framework that accurately measures disease-related risk factors in volumetric biomedical imaging, such as magnetic resonance imaging (MRI) scans, optical coherence tomography (OCT) scans, and ultrasound videos. To evaluate SLIViT, we applied it to five different datasets of these three different data modalities tackling seven learning tasks (including both classification and regression) and found that it consistently and significantly outperforms domain-specific state-of-the-art models, typically improving performance (ROC AUC or correlation) by 10-40%. Notably, compared to existing approaches, SLIViT can be applied even when only a small number of annotated training samples is available, which is often a constraint in medical applications. When trained on less than 700 annotated volumes, SLIViT obtained accuracy comparable to trained clinical specialists while reducing annotation time by a factor of 5,000 demonstrating its utility to automate and expedite ongoing research and other practical clinical scenarios.

Main

Biomedical imaging analysis is a critical component of clinical care with widespread use across multiple domains. For example, analyzing 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 and ultrasound. Monitoring heart function can help cardiologists assess potential cardiac issues, prescribe medications to improve a medical condition, e.g., 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.

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 and medical-related vision tasks in particular\(^7-9\). In order 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 annotated training samples are needed. Transfer learning addresses this challenge by pre-training a vision model for a
general learning task on a very large data set, and then using this general model as a starting point for training a specialized model on a much smaller data set\textsuperscript{10}. The key advantage of transfer learning is that the pre-training can be done on a large dataset in another domain, where annotated data are abundant, and then the fine-tuning can be done using a small dataset in the domain of interest. Using a transfer learning approach, a plethora of previously developed deep-vision models for 2D medical-imaging analysis\textsuperscript{11–14}, were first pre-trained on over a million labeled natural images taken from ImageNet\textsuperscript{15}, and then, fine-tuned to a specific medical-learning task on a much smaller number of biomedical images (typically fewer than 10,000). The availability of the labeled ImageNet dataset and the understanding that pre-trained weights can be leveraged as ‘prior knowledge’ for fine-tuning other learning tasks, were major factors in the fruitfulness of these 2D medical-imaging deep-vision models.

Many diagnoses rely, however, on volumetric biomedical imaging (e.g., 3D OCT and MRI scans, or ultrasound videos) and transfer learning is not directly applicable, since 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 labor-prohibitive than 2D images. For example, a 3D OCT scan that is composed of 97 2D slices (usually referred to as B-scans) usually requires a 5-10 minutes inspection of a highly trained clinical retina specialist in order to detect retinal-disease biomarkers, such as, the volume of a drusen lesion\textsuperscript{16}. 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 transfer learning. These gaps are acute because state-of-the-art models for 3D image analysis, such as 3D ResNet\textsuperscript{17}, involve the optimization of a very large number of parameters, thus requiring large annotated datasets for training.

Nonetheless, several attempts were undertaken to tackle volumetric medical-imaging 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\textsuperscript{18}. EchoNet was designed to predict heart ejection fraction (EF) in echocardiograms\textsuperscript{19}. In another recent study, a standard 3D ResNet was used by Witowski et al. to diagnose breast cancer in 3D MRI scans\textsuperscript{20}. The main limitation of each of these approaches is that they are all tailored and optimized for specific medical data 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.

Here, we present the SLice Integration by Vision Transformers (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 pre-trained 2D-based feature extractor and a vision transformer architecture. The 2D-based feature extractor allows leveraging prior 2D medical (and non-medical) vision knowledge when extracting information from each 2D frame of the volume. The attention-based mechanism of the transformer encoder allows next to integrate the extracted information across the 2D frames of the volume in question.

Specifically, we demonstrate the generalizability and utility of SLiViT in very different medical domains and data modalities, including retinal-disease risk biomarkers diagnosis in 3D OCT scans, cardiac function in echocardiogram videos, and hepatic disease severity assessment from 3D MRI scans. We show that SLiViT consistently attains significantly improved performance compared to domain-specific state-of-the-art models. Notably, the architecture stays invariant across data modalities, that is, SLiViT provides these improved performance results across data modalities without tailoring the architecture per data modality. We further 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 reduce the burden on clinicians and expedite ongoing research. Finally, we demonstrate that SLiViT is robust to frames permutation, and thus could be applied to datasets in which the slices order (within a volume) is not recorded, a recurring situation in currently available public limited datasets.

Results

A deep-vision model for analyzing volumetric medical data

In this study, we devise a deep learning model for automatic annotation of medical features in three-dimensional images (SLiViT). An overview of SLiViT is summarized in Figure 1. SLiViT preprocesses volumes into 2D images and then combines two deep-vision architectures: (1) a ConvNeXt backbone module that extracts feature maps for the slices (i.e., 2D frames of a volume), and (2) a Vision Transformer module that integrates the slices feature maps into a single diagnosis prediction. One key part of SLiViT is that its feature extractor is initialized by pre-trained weights. These weights were obtained by training a 2D ConvNeXt (T variant) first on ImageNet and then on an independent B-scan dataset to classify retinal-disease coarse risk factors. These pre-trained weights, that were used for initialization on each of the experiments detailed in this study, allowed SLiViT to improve the performance in a variety of learning tasks especially when a very small training dataset is available (few hundreds of samples). Our hypothesis was that the basic features that are extracted from B-scans when learning one task could serve as an improved training starting point not only for 3D OCT scans but
also for other data types, such as ultrasound video or 3D MRI, as they all share a basic set of features. In order to cope with volumetric data, which is essentially an array of 2D images, SLIViT tiles the 2D images into one elongated 2D image such that it conforms with the input dimension expected by the 2D-based feature extractor. Essentially, each original slice of the volume is embedded into a single feature map. Once the feature maps are extracted, they are comprehensively aggregated using a downstream vision transformer that aims to reconstruct the spatial signal that is lost when the volume is tiled into an elongated 2D image\textsuperscript{22,24}. Of note, the transformer’s attention mechanism implicitly eliminates the necessity for image registration preprocessing.

We tested SLIViT on five datasets of three different data modalities (OCT, ultrasound, and MRI) with a limited number of annotated samples, tackling a variety of learning tasks (including both classification and regression). Figure 2 summarizes the main results across these experiments, manifesting SLIViT’s consistent and significant superiority in performance compared to the corresponding (per-task) state-of-the-art method. In the OCT experiment, we evaluated the performance of an ocular disease high-risk factor diagnosis of SLIViT versus SLIVER-net\textsuperscript{18} 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 predictions’ $R^2$ of SLIViT respectively versus EchoNet\textsuperscript{19} in cardiac function analysis and versus 3D ResNet\textsuperscript{17,20} in hepatic fat levels imputation. In the following sections we present these and additional results in detail.

**SLIViT outperforms state-of-the-art models in detecting ocular disease high-risk factors using 3D OCT scans**

We first compared SLIViT’s performance against SLIVER-net and 3D ResNet on the Houston Dataset which includes only 691 OCT B-scan volumes of different individuals (see Methods). OCT B-scan volume data were collected from independent individuals affected in at least one eye by dry AMD, a globally leading cause of irreversible central visual impairment\textsuperscript{25}. Each OCT volume had four different binary labels of AMD high-risk biomarkers- drusen volume larger than 0.03 mm$^3$ (DV), intraretinal hyperreflective foci (IHRF), subretinal drusen deposits (SDD), and hyporeflective drusen cores (hDC)\textsuperscript{26}. The annotation was done by a senior retina specialist and the procured positive label frequencies of DV, IHRF, SDD, and hDC, were 47%, 43.5%, 52.8%, and 31.3%, respectively. 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 and PR AUCs (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 (see left panels
of Fig. 3, Fig. S1, and Table S1). 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’s ROC AUC was significantly better with a value of 0.924 while SLIVER-net and 3D ResNet achieved values of 0.838 and 0.777, respectively. In terms of average precision, SLIViT’s PR AUC for DV was significantly better with a value of 0.914 while SLIVER-net and 3D ResNet achieved values of 0.708 and 0.759, respectively.

To further challenge SLIViT we sought to explore its performance on the SLIVER-net Dataset used in the original SLIVER-net study. In this task, SLIVER-net should have an advantage as it was optimized for this dataset. The SLIVER-net Dataset was composed of roughly one thousand OCT scans (imaged from independent individuals in an Amish population) collected from three different clinical centers (see Methods). We trained SLIViT, SLIVER-net, and 3D ResNet 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 to 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 (see right panels of Figures 3 and S1, and Table S1).

SLIViT outperforms state-of-the-art models in analyzing cardiac function using ultrasound videos

In order 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 associated with a continuous number representing the corresponding ejection fraction (EF) measured in a clinical setting. The EF is measured by tracing the chamber volume of the left ventricle in the end-systole and end-diastole, and is a key metric of cardiac function as it measures how well the heart’s left ventricle is pumping blood. Low EF measurements (<0.5) can indicate cardiomyopathy or other heart problems. 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 EF measurements accordingly (>=0.5 was considered as normal) and, using the original EchoNet-Dynamic Dataset split, trained SLIViT and 3D ResNet (Fig. 4, upper panel). SLIViT obtained ROC AUC of 0.913 (CI [0.901, 0.928]) and significantly overperformed 3D ResNet with 0.793 ROC AUC (CI [0.772, 0.814]).

In a second experiment, we sought to test SLIViT in a regression task. Previously, Ghorbani, et al., implemented EchoNet, which is a GoogLeNet-based architecture for
predicting the EF of a given echocardiogram video, and obtained a 0.5 $R^2$ on the EchoNet-Dynamic Dataset test set. 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]). A scatter plot of the actual-versus-predicted per trained model is shown in the middle panel of Fig. 4 (see also ultrasound experiment in Fig. 2). As we did in all other experiments, we also tested 3D ResNet but its performance transcended neither of the two previous models. Moreover, we also examined R(2+1)D, a factorized spatiotemporal ResNet architecture 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\textsuperscript{17}, but ended up with poor results (not shown).

This result, together with the exceptional magnitude of this public annotated dataset, further motivated us to examine the dynamics of the training set size and SLiViT’s performance (Fig. 4, lower panel). We randomly sampled size-decreasing subsets from the original training set and trained a SLiViT model per subset. Compared to EchoNet trained on the original training set (n=7,465), SLiViT had on-par performance using only 25% of these data (n=1,866) and significantly better performance even on half of them for predicting the EF of a given echocardiogram. These observations demonstrate SLiViT’s ability to appropriately learn spatiotemporal features using a sparsely labeled dataset.

\textbf{SLiViT outperforms a state-of-the-art model in predicting hepatic fat levels in 3D MRI scans}

We next sought to evaluate SLiViT ability to model 3D MRI data. We used a UK Biobank Dataset containing 3D hepatic MRI scans and their corresponding measurement for hepatic proton density fat fraction (PDFF) level. 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\textsuperscript{28–30}. The development of a quantitative measurement of fat has been instrumental in improving the diagnosis of various fatty-liver and diabetes-related diseases\textsuperscript{31–35}. We removed unlabeled scans and preprocessed the rest of the dataset to contain only a single scan per individual. In this experiment we compared SLiViT only to 3D ResNet as it plays a double role- both the general and domain-specific state-of-the-art method\textsuperscript{20}. We randomly split the dataset and trained both models to measure PDFF levels of a given 3D MRI. SLiViT reached 0.829 $R^2$ (CI [0.817, 0.847]) and significantly outperformed 3D ResNet that obtained 0.611 $R^2$ (CI [0.566, 0.644]) (see MRI experiment in Fig. 2).
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 senior retina specialists (we used a majority vote when there was no consensus). We asked seven junior clinicians to (independently) 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 Fig. S3 summarize respectively the true positive rate (TPR; also known as recall) vs. false positive rate (FPR; also known as false alarm rate) and the positive predictive value (PPV; also known as precision) vs. 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 seconds). Interestingly, SLIViT obtained considerably lower performance in the hDC classification task compared to 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 cases and normals.

SLIViT is robust to within-volume frames 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. Figure S4 shows the average bootstrapped ROC AUC dispersion of these 101 models. Interestingly, the original model did not outperform the shuffled models. We observed that compared to the 100 shuffled-models performance, the average rank of the original model across the four AMD biomarkers was 40. This finding suggested that even if the original order is not documented, SLIViT’s performance does not deteriorate.

Discussion

Procuring tens of thousands of annotated 3D medical-imaging samples to train standard 3D vision models is expert-time prohibitive, impeding the full optimization of such models. In this work we devised SLIViT, an AI-based framework that allows an accurate analysis
of potentially any 3D medical-imaging dataset. SLiViT leverages a unique combination of deep-vision modules and ‘prior knowledge’ from the 2D domain. This allows it to be adept at such medical-imaging-learning tasks, in which the number of annotated training samples is typically very limited, and significantly outperform 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 medical domains (retinal, cardiac, and hepatic) across different 3D medical-imaging data modalities (OCT, echocardiograms, and MRI) against domain-specific 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 labeled with a corresponding hepatic fat content measurement and again, observed significant improvement compared to the state-of-the-art. We also showed that SLiViT was able to obtain on-par performance to clinical specialists’ assessment, but rather, four orders of magnitude faster compared to the annotation procurement net time required by the specialists. Lastly, 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 potentially agnostic to imaging protocol.

We hypothesize that SLiViT may be further improved by applying task-specific changes to the model’s architecture. Specifically, in all the experiments done in this study, the same architecture was used and the weights were initialized similarly, where the only difference was the per-task fine-tuning. In its current form, SLiViT can serve as a reliable baseline model for any study of volumetric biomedical imaging.

SLiViT was tested on 3D OCT scans, echocardiograms, and MRI volumes and can potentially be leveraged to analyze other types of data modalities, such as 3D computed tomography (CT) scans and 3D X-ray imaging. Such medical volumetric imaging data is 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 analyzing a series of medical frames created in a structured medical-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, objects may flip, change color (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-videos-learning tasks (such as action recognition). That being said, SLIViT could potentially be tweaked to perform well on natural videos as well, e.g., using a different feature extractor, however, this direction requires further research.

Importantly, there are multiple additional steps that are required in order to deploy SLIViT in a clinical setting. Notably, the point of operation (tradeoff 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 (see Fig. 5 and Fig. S3). Moreover, additional evaluations of the models are required to ensure no systematic biases exist that would lead to increasing health disparities.

Overall, this study highlights an important step toward 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 medical-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 will always exist. Once a previously unknown disease-related risk factor is found and characterized, it could take months in order 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.

Methods

SLIViT’s development and analysis

SLIViT was implemented in Python using PyTorch\textsuperscript{36} v1.10.2, fast.ai\textsuperscript{37} v2.6.3, and scikit-learn\textsuperscript{38} v1.0.2 libraries (full libraries and version list can be found at \url{https://github.com/berkindurmus/SLIViT/blob/main/requirements.txt}).

Model specifications

The SLIViT framework contains a preprocessing step, a 2D ConvNeXt that serves as a feature extractor, and a vision transformer (ViT) that serves as a feature integrator (see
A ConvNeXt architecture has several complexities. Here we used the backbone of the tiny variant (ConvNeXt-T) with 256x256 image size as SLIViT’s feature extractor. In the ViT we applied a few adjustments with respect to the original architecture, including using GeLu as the activation functions and initializing the positional embeddings as the number of the original slice.

Let N be the number of HxW 2D frames of an input image. Given an input WxHxN image, its N frames are resized (according to the ConvNeXt-T variant) and tiled into an image of size N*256x256 (see Step (1) in Fig. 1). The manipulated image is then fed into the feature extractor which generates, in turn, an N*8x8 feature maps with F=768 filters each. These feature maps are then reshaped into N different 8x8x768 feature maps (see Step (3) in Fig. 1), each corresponding to a slice in the original volume. Each of the feature maps is flattened into an 8*8*768 (1D) vector and projected into a vector of size 768 using a fully connected (FC) layer. Of note, the bias term of the FC layer is initialized as the feature map number (that essentially corresponds to an original slice number). The projected feature volumes 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 FC layer to generate final output.

Pre-training

We borrowed an ImageNet-1K pre-trained SLIViT-like feature extractor architecture, i.e., a ConvNeXt-T backbone, from https://huggingface.co/facebook/convnext-tiny-224, and appended to it a subsequent FC layer to fit a four-category classification task. We then further trained this SLIViT backbone-like module on the publicly available Kermany Dataset. Of note, in each of the experiments done in this study, SLIViT was initialized using these pre-trained backbone weights and fine-tuned per task. The pre-trained backbone weights are available at project’s GitHub repository (see Code Availability section).

Per-task fine-tuning

All the experiments done in this work were initialized with the abovementioned backbone weights and then were fine-tuned (as a whole) according to the dataset and task in question. We applied standard preprocessing transformations (as implemented in PyTorch) on every training set such as contrast stretching. We used the binary cross entropy and L1 norm as loss functions for the classification and regression tasks, respectively. In each experiment, a random validation set was used for determining the convergence of the training with the same loss function metric used for the test set evaluation. 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\textsuperscript{41,42}. 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 chosen for the test set evaluation. We used Weights & Biases for experiment tracking and visualizations of the training procedures\textsuperscript{43}.

**Statistical Analysis**

The performance of each trained model was evaluated (on the corresponding test set) using an appropriate metric score. The binary classification tasks were evaluated using 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\% confidence interval (CI) was computed for each evaluated score using a standard bootstrapping procedure with 1,000 iterations as done in other studies\textsuperscript{44}. Briefly, let $n$ denote the test set size, for each bootstrap iteration $n$ samples were randomly drawn (with repetition) and a single score based on the predictions of the sampled set was obtained. Out of the 1,000 sampled-sets scores distribution, the 50th and 950th ranked scores were selected to obtain the 90\% CI. In order 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 corresponding sampled-set scores was computed ($H_0: \mu \neq 0$). SLiViT’s performance improvement was concluded as significant if the t-test produced a P-value lower than 1e-7.

**Datasets**

**The Houston Dataset**

1,128 patients were diagnosed with intermediate AMD in their scanned eye by clinical examination (Beckman Classification\textsuperscript{45}) at the Retina Consultants of Texas Eye Clinics 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, Inc, Heidelberg, Germany) 6x6 mm (fovea centered, 10X10 degrees; 49 B-scans spaced 122 microns apart, ART=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.
The AMD-biomarker analysis was conducted at the Doheny Image Reading Research Laboratory (DIRRL) in compliance with the Declaration of Helsinki and approved by the 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) 48.23% of the scans had drusen volume > 0.03 mm$^3$ within the 3 central mm$^2$ (denoted DV); (2) 36.17% of the scans had intraretinal hyperreflective foci (denoted IHRF); (3) 31.45% of the scans had subretinal drusenoid deposits (SDD); and (4) 11.27% of the scans had hyporeflective drusen core (hDC). Of note, some scans were positive for more than one biomarker.

The SLIVER-net Dataset

The SLIVER-net Dataset, which was originally used by Rakocz and others$^{18}$ to tune and validate SLIVER-net, was collected from three independent medical centers between February 2013 and July 2016$^{18}$. The dataset consisted of 1,007 OCT volumes each consisting of 97 B-scans (97,679 B-scans overall) collected from 649 subjects 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 centers in Pennsylvania, Indiana, and Ohio under the supervision of investigators at the University of Pennsylvania (UPEN), University of Miami (MU), and Case Western Reserve University (CWRU), respectively. All OCT B-scan volumes in this dataset were acquired with the Heidelberg Spectralis OCT using a scan pattern centered on the fovea (20°x20°; 97 B-scans; 512 A-scans per B-scans; ART 9). In order 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 to 49 B-scans. Also, to avoid aliasing, we applied an anti-aliasing filter on OCT volumes.

The positive label frequencies in this dataset were 3.37%, 7.87%, 2.0%, and 2.67%, 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 3D OCT B-scan volumes (fovea centered, 10x10 degree, ART=5) collected from 205 individuals at the Doheny-UCLA Eye Centers in Pasadena between 2013 and 2022. Each of the OCT volumes was acquired on the Heidelberg Spectralsis HRA+Optical Coherence Tomography (OCT SPECTRALIS; Heidelberg Engineering, Inc, Heidelberg, Germany). Out of the 205 OCT volumes, 198 contained 97 B-scans and seven contained 49 B-scans. The OCT B-scans 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 32.8%, 51.6%, 42.9%, and 12.5%, for DV, IHRF, SDD, and hDC, respectively.

The EchoNet-Dynamic Dataset

The EchoNet-Dynamic Dataset\(^27\) was downloaded on September 7, 2022 from https://echonet.github.io/dynamic/index.html#dataset. The dataset contains 10,030 echocardiograms (heartbeat ultrasound videos) obtained from 10,030 different individuals who underwent echocardiography between 2006 and 2018. Each echocardiogram was labeled with a continuous number (between zero and one) representing the ejection fraction (EF). The EF was obtained by a registered sonographer and further verified by a level 3 echocardiographer. The minimal EF in the dataset was 0.069 while the maximal was 0.97. The average EF 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 (2D images) 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.

The United Kingdom Biobank Dataset

The United Kingdom Biobank (UKBB) Dataset of MRI imaging with Proton Density Fat Fraction (PDFF) measurements was downloaded on June 7, 2022, from the UKBB\(^49\). 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 5972 (60%), 1991 (20%), and 1991 (20%), respectively.

Code availability

The code of SLIViT is available at the project’s GitHub repository:
https://github.com/berkindurmus/SLIViT.

Acknowledgments

This work was supported by NIH/NEI grants RO1EY023164 and 1R01EY030614 and an
Unrestricted Grant from Research to Prevent Blindness, Inc. This research was
carried out using the UK Biobank Resource under application #33127.
References


43. Weights & Biases – Developer tools for ML. Available at: https://wandb.ai/site. (Accessed: 16th May 2023)


Figure 1 | The proposed SLIViT framework

The input of SLIViT is a 3D volume of N frames of size HxW. (1) The frames of the volume are resized and vertically tiled into an "elongated image". (2) The elongated image is fed into the pre-trained 2D Feature Extractor. (3) N feature maps are extracted (each corresponds to an original frame). (4) Feature maps are (flattened and) fed into a multi-head attention transformer encoder followed by a fully-connected layer that outputs the prediction for the task in question.
Figure 2 | SLIViT outperforms domain-specific state-of-the-art (SOTA) methods on different learning tasks and 3D medical-imaging data modalities.

Shown are the performance scores in one classification task (with two different metrics) of eye disease biomarker diagnosis in volumetric OCT scans and two regression tasks of (1) heart function analysis in ultrasound videos and (2) liver fat levels imputation in volumetric MRI scans. Box plots whiskers extend to the 5th and the 95th percentiles (of the bootstrapped scores distribution).
Figure 3 | ROC AUC performance comparison of three models in four independent AMD-biomarker classification tasks when trained on less than 700 OCT volumes. Shown are the ROC AUCs of SLiViT (blue), SLIVER-net (orange), and 3D ResNet (green) on single-prediction tasks of four AMD high-risk factors in two independent OCT volume datasets. The left panel shows the performance when trained and tested on the Houston Dataset. The right panel shows the performance when trained on the Houston Dataset and tested on the SLIVER-net Dataset. Each box was obtained by bootstrapping the test scores for each model in the corresponding classification task (see Table S1A). Box plots whiskers extend to the 5th and the 95th percentiles.
Figure 4 | Performance comparison on cardiac function prediction tasks using echocardiograms

Upper panel - ROC curves of cardiomyopathy prediction (EF<0.5). Middle panel - predicted vs. actual EF levels for three different models trained on the original training set (solid black line represents the y=x line). Lower panel - $R^2$ performance of heart EF prediction (a regression task) using different percentages of the original training dataset. Each box was obtained by bootstrapping the test scores of the corresponding model. Box plots whiskers extend to the 5th and the 95th percentiles. Of note, when SLIViT was trained on 25% (n=1,866) of the original training set it obtained similar accuracy as EchoNet when trained on 100% (n=7,465) of the training set.
Figure 5 | SLIVIT’s ROC curve compared to junior clinical retina specialists’ assessment

Shown are the ROC curves (blue) of SLIVIT trained to predict four AMD high-risk biomarkers (DV, IHRF, SDD, and hDC; see main text) using less than 700 OCT volumes (Houston Dataset) and tested on an independent dataset (Pasadena Dataset). The light-blue shaded area represents the 90% confidence interval for SLIVIT’s performance. The red dot represents the specialists’ average performance. The green asterisks correspond to the retina specialists’ assessments. Two of the clinical specialists obtained the exact same performance score for IHRF classification.

Area Under Curve: 0.907

Area Under Curve: 0.858

Area Under Curve: 0.87

Area Under Curve: 0.709
Supplementary Material

Figure S1 | PR-AUC performance comparison of three models in four independent AMD-biomarker classification tasks when trained on less than 700 OCT volumes.

Shown are the PR AUCs as an alternative scoring metric for the experiment shown in Figure 3. The left panel shows the performance when trained and tested on the Houston Dataset. The right panel shows the performance when trained on the Houston Dataset and tested on the SLIVER-net Dataset. Each box was obtained by bootstrapping the test scores for each model in the corresponding classification task (see Table S1B). Box plots whiskers extend to the 5th and the 95th percentiles. The dashed lines represent the corresponding positive label prevalence, and thus the expected performance of a naive classifier for the corresponding biomarker.
Figure S2 | Performance comparison of a cardiomyopathy binary classification task on echocardiograms

Shown are the PR curves yielded by modeling SLIViT (blue) and 3D ResNet (green) to classify cardiomyopathy.
Figure S3 | SLiViT’s PR performance compared to junior clinical retina specialists’ assessment

Shown are the PR curves (blue) of SLiViT trained to predict four AMD high-risk biomarkers (DV, IHRF, SDD, and hDC; see main text) using less than 700 OCT volumes (Houston Dataset) and tested on an independent dataset (Pasadena Dataset). The light-blue shaded area represents the 90% confidence interval for SLiViT’s performance. The red dot represents the specialists’ average performance. The green asterisks correspond to the retina specialists’ assessments. Two of the clinical specialists obtained the exact same performance score for IHRF classification.
Figure S4 | SLiViT performance in a classification task using shuffled OCT volumes

Shown is the ROC AUC distribution of 100 SLiViT models trained on 100 different shuffled copies of an OCT volume dataset (light blue). Box plots whiskers extend to the 5th and the 95th percentiles. The dashed blue line represents the ROC AUC of a SLiViT model of the OCT volume dataset using the original order of each volume. The performance ranks of this latter model compared to the former models’ distribution were 22, 34, 56, and 47 for DV, IHRF, SDD, and hDC, respectively.
Shown are the underlying performance raw numbers of Fig. 3 (ROC AUC) and Fig. S1 (PR AUC) of the AMD high-risk biomarker prediction experiments. The numbers in the square brackets represent the 90% confidence interval as obtained by bootstrapping the predictions (see Methods section for further details).

### A – ROC AUCs

<table>
<thead>
<tr>
<th>Test dataset</th>
<th>Method</th>
<th>DV</th>
<th>IHRF</th>
<th>SDD</th>
<th>hDC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Houston</td>
<td>SLIViT</td>
<td>.924</td>
<td>.883</td>
<td>.877</td>
<td>.89</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[.909, .938]</td>
<td>[.86, .906]</td>
<td>[.855, .893]</td>
<td>[.877, .916]</td>
</tr>
<tr>
<td></td>
<td>SLIVER-net</td>
<td>.838</td>
<td>.837</td>
<td>.805</td>
<td>.854</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[.813, .86]</td>
<td>[.82, .855]</td>
<td>[.78, .827]</td>
<td>[.836, .869]</td>
</tr>
<tr>
<td></td>
<td>3D ResNet</td>
<td>.777</td>
<td>.655</td>
<td>.783</td>
<td>.782</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[.769, .783]</td>
<td>[.625, .682]</td>
<td>[.762, .806]</td>
<td>[.757, .805]</td>
</tr>
</tbody>
</table>

### B – PR AUCs

<table>
<thead>
<tr>
<th>Test dataset</th>
<th>Method</th>
<th>DV</th>
<th>IHRF</th>
<th>SDD</th>
<th>hDC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Houston</td>
<td>SLIViT</td>
<td>.958</td>
<td>.891</td>
<td>.967</td>
<td>.863</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[.941, .975]</td>
<td>[.873, .909]</td>
<td>[.959, .973]</td>
<td>[.839, .892]</td>
</tr>
<tr>
<td></td>
<td>SLIVER-net</td>
<td>.933</td>
<td>.839</td>
<td>.911</td>
<td>.625</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[.919, .95]</td>
<td>[.817, .86]</td>
<td>[.9, .922]</td>
<td>[.576, .676]</td>
</tr>
<tr>
<td></td>
<td>3D ResNet</td>
<td>.904</td>
<td>.8</td>
<td>.895</td>
<td>.716</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[.891, .911]</td>
<td>[.788, .813]</td>
<td>[.865, .925]</td>
<td>[.689, .737]</td>
</tr>
<tr>
<td>SLIVER-net</td>
<td>SLIViT</td>
<td>.914</td>
<td>.852</td>
<td>.855</td>
<td>.795</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[.898, .928]</td>
<td>[.826, .875]</td>
<td>[.831, .879]</td>
<td>[.747, .838]</td>
</tr>
<tr>
<td></td>
<td>SLIVER-net</td>
<td>.708</td>
<td>.799</td>
<td>.785</td>
<td>.74</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[.676, .744]</td>
<td>[.778, .817]</td>
<td>[.752, .816]</td>
<td>[.716, .76]</td>
</tr>
<tr>
<td></td>
<td>3D ResNet</td>
<td>.759</td>
<td>.619</td>
<td>.791</td>
<td>.669</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[.748, .769]</td>
<td>[.584, .647]</td>
<td>[.77, .815]</td>
<td>[.622, .697]</td>
</tr>
<tr>
<td>SLIVER-net</td>
<td>SLIViT</td>
<td>.575</td>
<td>.728</td>
<td>.399</td>
<td>.222</td>
</tr>
<tr>
<td></td>
<td>[0.517, 0.63]</td>
<td>[0.696, 0.763]</td>
<td>[0.341, 0.469]</td>
<td>[0.184, 0.263]</td>
<td></td>
</tr>
<tr>
<td>-------------</td>
<td>---------------</td>
<td>---------------</td>
<td>---------------</td>
<td>---------------</td>
<td></td>
</tr>
<tr>
<td>SLIVER-net</td>
<td>0.535</td>
<td>0.621</td>
<td>0.278</td>
<td>0.093</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.47, 0.588]</td>
<td>[0.588, 0.653]</td>
<td>[0.221, 0.345]</td>
<td>[0.07, 0.122]</td>
<td></td>
</tr>
<tr>
<td>3D ResNet</td>
<td>0.497</td>
<td>0.593</td>
<td>0.183</td>
<td>0.219</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.444, 0.553]</td>
<td>[0.563, 0.626]</td>
<td>[0.147, 0.225]</td>
<td>[0.162, 0.282]</td>
<td></td>
</tr>
</tbody>
</table>
Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- nreditorialpolicychecklist.pdf