Comprehensive Evaluation of Similarity-Based Positional Encoding for Medical Vision Transformers

Anonymous authors
Paper under double-blind review

Abstract

Similarity-based positional encoding (SimPE) for Vision Transformers has been shown to improve classification accuracy on medical imaging benchmarks. However, accuracy alone is insufficient to characterise the practical value of a positional encoding strategy for clinical applications. This paper provides a comprehensive evaluation of SimPE addressing three open questions: (i) *how SimPE affects the structure of the learned feature representations*; (ii) *whether the accuracy advantage extends to other classification metrics*; (iii) *what is the computational cost of SimPE*. We show through UMAP dimensionality reduction that SimPE induces more discriminative class clusters than standard learned positional encoding (SLPE) and Rotary Position Embedding (RoPE) across all six tested modalities. Ten-fold cross-validation with 95% confidence intervals on accuracy, precision, recall, and F1 score confirms that SimPE outperforms alternative approaches consistently, with narrow confidence intervals indicating stable training dynamics. Per-epoch training time and total-convergence analysis show that SimPE is only marginally slower than SLPE per epoch but converges in fewer epochs, usually yielding shorter wall-clock training time; RoPE is substantially more expensive in both dimensions. Attention map analysis further confirms that SimPE focuses on clinically relevant image structures more reliably than the alternatives. Finally, SimPE extends naturally to volumetric (3D) medical images with consistent improvements. These results consolidate SimPE as a leading positional encoding strategy for geometry-structured medical image classification.

1 Introduction

Positional encoding (PE) is a foundational design choice in Vision Transformers (ViTs) (Dosovitskiy et al., 2021): it supplies the otherwise permutation-invariant self-attention mechanism with information about the spatial arrangement of image patches (Vaswani et al., 2017). In medical imaging, where pathologies often manifest as coherent geometric patterns, the form of this encoding can have a strong influence on what the model learns.

Leonardi et al. (2024) and Santomauro et al. (2026) introduced SimPE (Simnilarity-based Positional Encoding), a positional encoding mechanism that derives spatial information from pairwise cosine similarity between convolutional feature maps. Those works demonstrated that SimPE achieves consistently higher classification accuracy than the two main state-of-the-art approaches: standard learned PE (SLPE) (Dosovitskiy et al., 2021) and Rotary Position Embedding (RoPE) (Su et al., 2023), across six MedMNIST datasets (Yang et al., 2021; 2023) under both from-scratch and fine-tuning regimes.

The present paper extends that evaluation along three dimensions critical for clinical practice:

Representation quality. Accuracy on a held-out benchmark reflects the decision boundary quality, not the underlying structure of learned representations. A PE that produces well-separated class clusters is likely to generalise better under distribution shift—a critical concern for medical AI.

Multi-metric performance. Medical classification is rarely evaluated on accuracy alone. Precision, recall, and F1 carry distinct clinical implications, and confidence intervals over cross-validation folds are essential for reliable comparison.

Training efficiency. Deployment pipelines face computational constraints. A method requiring significantly more training time may be impractical even if nominally more accurate.

The primary contributions are: (i) a multi-metric 10-fold cross-validated evaluation with 95% CIs; (ii) a training efficiency analysis; (iii) attention map analysis confirming clinically relevant focus; (iv) UMAP representation quality analysis across all six datasets; and (v) an extension of SimPE to volumetric 3D images. The paper is organized as follows: Section 2 surveys positional encoding methods; Section 3 describes the SimPE method and the experimental setup. Sections 4–7 report multi-metric performance, training efficiency, attention maps, and representation quality; Section 8 presents the 3D extension, and Section 9 provides the conclusions.

2 Positional Encoding: Related Work

Positional encoding methods (Jiang et al., 2022) provide the permutation-invariant self-attention (Vaswani et al., 2017) with spatial context. In the context of ViTs they fall into five categories: [1] (a) *Fixed*: sinusoidal functions of absolute position (Vaswani et al., 2017); (b) *Standard Learned (SLPE)*: a learnable matrix optimised during training (Dosovitskiy et al., 2021); (c) *Learned Relative*: focuses on relative distances, useful for segmentation (Wu et al., 2021); (d) *RoPE*: incorporates relative position via rotation of query and key vectors (Su et al., 2023), requiring architectural changes incompatible with standard fine-tuning; 2D extensions for medical ViTs exist (Li et al., 2025); (e) *Task-specific*: methods tailored to particular modalities, e.g., volumetric PE for 3D CT/MRI (Wang et al., 2022; Tang et al., 2022; Wang et al., 2021) or circular encodings for 3D segmentation (Yu & Triesch, 2023).

Task-specific methods are non-generalizable, and learned relative PE applies a clipping that limits long-range relationships. For general ViT classification, the relevant baselines are therefore SLPE and RoPE, against which we compare SimPE in all experiments.

3 Method and Experimental Setup

SimPE. Let I be the input image partitioned into N patches. A standard linear projection yields patch embeddings $X \in \mathbb{R}^{N\times d}$. A shallow convolutional network applied to I produces a geometric feature map $F \in \mathbb{R}^{N\times d}$. The pairwise cosine similarity between all patch feature vectors is assembled into $S \in \mathbb{R}^{N\times N}$:

$$S = \frac{F\,F^{\top}}{\|F\|_2}, \tag{1}$$

where $\|\cdot\|_2$ normalises the rows of F. A learnable projection $W \in \mathbb{R}^{N\times d}$ maps S to the embedding dimension and the result is added to the patch embeddings:

$$PE = SW, \qquad X' = X + PE. \tag{2}$$

The modified embedding X' feeds the standard transformer encoder without any architectural change. Unlike RoPE, SimPE does not modify the attention computation and is therefore compatible with any pre-trained ViT. The key property is that SimPE encodes *what* patches look like relative to one another rather than *where* they are in the grid, providing an inductive bias aligned with how pathological features are distributed in medical images.

[1] We use the terms *positional encoding* for the method and *positional embedding* for the numerical vector it produces (Irani & Metsis, 2025).

Table 1: Architectural and training hyperparameters.

Config	Hyperparameter	Value
Architecture	Transformer Layers	4 (scratch), 12 (fine-tuned)
	Attention Heads	4 (scratch), 12 (fine-tuned)
	Embedding Size	768
	MLP Dimension	256 (scratch), 3072 (fine-tuned)
Training	Optimizer	Adam
	Learning Rate	10^{-4} (scratch), 10^{-3} (fine-tuned)
	Scheduler	CosineAnnealingLR
	Batch Size	256 (scratch), 512 (fine-tuned)
	Input Image Size	224×224
	Gradient Clipping	1.0
	RGB Normalisation (mean)	[0.485, 0.456, 0.406]

Datasets. Six 2D classification tasks from MedMNIST (Yang et al., 2021; 2023): **BloodMNIST** (8-class microscopy (Acevedo & Merino, 2020)), **PneumoniaMNIST** (binary X-ray (Wang et al., 2017)), **PathMNIST** (9-class colorectal pathology (Bejnordi et al., 2017)), **DermaMNIST** (7-class dermoscopy (Tschandl, 2018)), **BreastMNIST** (binary ultrasound (Al-Dhabyani et al., 2020)), **TissueMNIST** (8-class histology (Vebjorn et al., 2012)). All images are used at the 224×224 resolution provided by the MedMNIST dataset creators.

Models and training. Two ViT variants: a *scratch* model (4 layers, 4 heads, dim 768, MLP 256) and a *fine-tuned* model (12 layers, 12 heads, MLP 3072) initialised from an ImageNet (Deng et al., 2009) pretrained checkpoint. Both are trained with Adam (Kingma & Ba, 2014), cosine annealing LR (Loshchilov & Hutter, 2017), and early stopping on validation accuracy. All runs use a single NVIDIA A100 GPU. Table 1 summarises the hyperparameter configuration.

Evaluation. Classification metrics (accuracy, precision, recall, F1) are reported as mean ± 95% CI over 10-fold cross-validation. Training efficiency is measured as per-epoch wall-clock time and total training time until early stopping. Representation quality is assessed via UMAP (McInnes et al., 2018) projections of test-set `[CLS]` token activations (15 neighbours for small datasets, 40 for large).

4 Multi-Metric Performance

Tables 2 and 3 report 10-fold cross-validation results for the scratch-trained and fine-tuned models.

Scratch model. When trained from scratch, SimPE achieves the best F1 on all six datasets and the best accuracy on four of six. The largest gains over SLPE are on BloodMNIST (+8.7pp), PneumoniaMNIST (+8.7pp), and TissueMNIST (+8.2pp)—datasets where spatial geometry is most discriminative. These results suggest that SimPE provides a more effective spatial inductive bias precisely when no external pretraining is available to compensate for a weak positional signal.

RoPE shows markedly less stable behaviour in this regime. Although it yields marginal accuracy gains on BreastMNIST and PneumoniaMNIST over SLPE, it severely degrades on PathMNIST (accuracy 85.2% → 67.0%, F1 0.86 → 0.65) and shows substantial F1 drops on BreastMNIST, DermaMNIST, and TissueMNIST. This instability suggests that RoPE's rotation-based positional signal interacts poorly with the texture statistics of certain histological and dermoscopic images when no pretrained initialisation is provided.

Table 2: Performance comparison of positional encoding strategies. Accuracy in percentage, F1 as a score. Values: mean ± 95% CI. Best per row in **bold**.

Model	Dataset	Accuracy (%)			F1		
		SLPE	RoPE	SimPE	SLPE	RoPE	SimPE
Fine-tuned	BloodMNIST	87.4 ± 0.8	95.6 ± 0.5	$\mathbf{96.4 \pm 0.3}$	0.84 ± 0.01	0.95 ± 0.01	$\mathbf{0.96 \pm 0.00}$
	BreastMNIST	73.2 ± 4.7	$\mathbf{85.0 \pm 2.9}$	83.0 ± 2.5	0.83 ± 0.05	0.78 ± 0.04	$\mathbf{0.89 \pm 0.04}$
	DermaMNIST	68.1 ± 1.4	77.7 ± 0.9	$\mathbf{79.0 \pm 1.1}$	0.25 ± 0.02	0.53 ± 0.01	$\mathbf{0.58 \pm 0.02}$
	PathMNIST	94.0 ± 1.6	97.0 ± 0.4	$\mathbf{97.9 \pm 0.2}$	0.94 ± 0.00	0.97 ± 0.01	$\mathbf{0.98 \pm 0.00}$
	PneumoniaMNIST	85.3 ± 2.3	95.6 ± 0.6	$\mathbf{95.8 \pm 0.6}$	0.89 ± 0.02	0.95 ± 0.01	$\mathbf{0.96 \pm 0.01}$
	TissueMNIST	55.4 ± 0.7	$\mathbf{62.7 \pm 0.3}$	61.1 ± 0.3	0.38 ± 0.01	$\mathbf{0.50 \pm 0.00}$	0.48 ± 0.01
Scratch	BloodMNIST	74.4 ± 2.6	75.6 ± 1.0	$\mathbf{83.1 \pm 2.5}$	0.76 ± 0.02	0.70 ± 0.08	$\mathbf{0.82 \pm 0.02}$
	BreastMNIST	73.1 ± 4.1	$\mathbf{74.8 \pm 2.6}$	73.1 ± 4.4	0.83 ± 0.04	0.56 ± 0.08	$\mathbf{0.84 \pm 0.01}$
	DermaMNIST	68.3 ± 1.4	67.2 ± 0.8	$\mathbf{70.0 \pm 1.0}$	0.30 ± 0.02	0.19 ± 0.01	$\mathbf{0.39 \pm 0.01}$
	PathMNIST	85.2 ± 1.6	67.0 ± 0.5	$\mathbf{88.0 \pm 1.1}$	0.86 ± 0.01	0.65 ± 0.01	$\mathbf{0.88 \pm 0.01}$
	PneumoniaMNIST	85.2 ± 1.6	87.4 ± 0.9	$\mathbf{93.9 \pm 0.6}$	0.87 ± 0.04	0.83 ± 0.01	$\mathbf{0.95 \pm 0.01}$
	TissueMNIST	43.1 ± 1.4	43.1 ± 1.4	$\mathbf{51.3 \pm 0.6}$	0.32 ± 0.01	0.23 ± 0.03	$\mathbf{0.37 \pm 0.01}$

Table 3: Precision and Recall (mean ± 95% CI). Best per row in **bold**.

Model	Dataset	Precision			Recall		
		SLPE	RoPE	SimPE	SLPE	RoPE	SimPE
Fine-tuned	BloodMNIST	0.85 ± 0.02	0.95 ± 0.01	$\mathbf{0.96 \pm 0.00}$	0.84 ± 0.01	0.95 ± 0.01	$\mathbf{0.96 \pm 0.00}$
	BreastMNIST	0.73 ± 0.07	0.80 ± 0.05	$\mathbf{0.84 \pm 0.05}$	$\mathbf{0.97 \pm 0.04}$	0.76 ± 0.07	0.94 ± 0.05
	DermaMNIST	0.25 ± 0.04	0.55 ± 0.02	$\mathbf{0.60 \pm 0.02}$	0.25 ± 0.02	0.52 ± 0.02	$\mathbf{0.56 \pm 0.03}$
	PathMNIST	0.94 ± 0.00	0.97 ± 0.01	$\mathbf{0.98 \pm 0.00}$	0.95 ± 0.00	0.97 ± 0.01	$\mathbf{0.98 \pm 0.00}$
	PneumoniaMNIST	0.85 ± 0.03	0.95 ± 0.01	$\mathbf{0.96 \pm 0.01}$	0.94 ± 0.02	0.95 ± 0.01	$\mathbf{0.97 \pm 0.00}$
	TissueMNIST	0.39 ± 0.01	$\mathbf{0.51 \pm 0.00}$	0.49 ± 0.01	0.38 ± 0.01	$\mathbf{0.50 \pm 0.00}$	0.47 ± 0.01
Scratch	BloodMNIST	0.79 ± 0.03	0.71 ± 0.17	$\mathbf{0.84 \pm 0.30}$	0.73 ± 0.03	0.70 ± 0.02	$\mathbf{0.80 \pm 0.03}$
	BreastMNIST	0.71 ± 0.06	0.57 ± 0.15	$\mathbf{0.73 \pm 0.02}$	$\mathbf{1.00 \pm 0.00}$	0.56 ± 0.07	$\mathbf{1.00 \pm 0.00}$
	DermaMNIST	0.31 ± 0.04	0.17 ± 0.02	$\mathbf{0.40 \pm 0.01}$	0.30 ± 0.03	0.22 ± 0.01	$\mathbf{0.38 \pm 0.02}$
	PathMNIST	0.86 ± 0.01	0.65 ± 0.01	$\mathbf{0.89 \pm 0.01}$	0.86 ± 0.01	0.65 ± 0.01	$\mathbf{0.88 \pm 0.01}$
	PneumoniaMNIST	0.95 ± 0.05	0.85 ± 0.01	$\mathbf{0.96 \pm 0.01}$	0.81 ± 0.05	0.82 ± 0.01	$\mathbf{0.94 \pm 0.01}$
	TissueMNIST	0.32 ± 0.02	0.21 ± 0.04	$\mathbf{0.38 \pm 0.01}$	0.32 ± 0.02	0.25 ± 0.02	$\mathbf{0.36 \pm 0.01}$

Fine-tuned model. Fine-tuning from ImageNet weights improves all three strategies, but SimPE remains the top performer overall. It achieves the best F1 on 5 of 6 datasets and near-saturation performance on BloodMNIST, PathMNIST, and PneumoniaMNIST (F1: 0.96, 0.98, 0.96). These results show that SimPE's geometry-aware encoding remains beneficial even when strong visual representations are transferred from large-scale pre-training.

RoPE is more competitive in the fine-tuned setting, achieving the highest accuracy on BreastMNIST (85.0%) and TissueMNIST (62.7%). However, this does not translate to the best F1. On BreastMNIST, RoPE reaches the highest accuracy but its F1 (0.78) is lower than both SLPE (0.83) and SimPE (0.89), indicating an unfavourable precision–recall balance. On TissueMNIST, RoPE leads in both accuracy and F1, although the margin over SimPE is modest (62.7% vs. 61.1% accuracy; F1 0.50 vs. 0.48).

Precision, recall, and confidence interval width. Table 3 expands the picture with precision and recall, and the width of the confidence intervals reveals training stability differences that accuracy alone conceals. SimPE achieves the best precision and recall on the majority of dataset–regime combinations. On DermaMNIST scratch its F1 CI is ±0.01, versus ±0.02 for SLPE; on BloodMNIST scratch SimPE achieves ±2.5pp in accuracy versus ±1.0pp for RoPE, which occasionally experiences fold-level collapse (large CI on precision). Narrower CIs indicate that SimPE's advantage is structural: it is unlikely to degrade on an unseen fold.

Table 4: Average training time per epoch (mm:ss). Δ_{SLPE}: SimPE overhead vs. SLPE (s). Δ_{RoPE}: SimPE saving vs. RoPE (s; negative = SimPE faster).

Paradigm	Dataset	SLPE	RoPE	SimPE	Δ_{SLPE}	Δ_{RoPE}
Fine-tuning	BloodMNIST	00:52	01:23	00:55	+3 s	−28 s
	BreastMNIST	00:03	00:02	00:03	0 s	+1 s
	DermaMNIST	00:32	01:07	00:35	+3 s	−32 s
	PathMNIST	06:52	07:26	07:02	+10 s	−24 s
	PneumoniaMNIST	00:21	00:55	00:22	+1 s	−33 s
	TissueMNIST	11:57	10:36	12:44	+47 s	+128 s
From scratch	BloodMNIST	03:49	01:43	01:46	+3 s	+3 s
	BreastMNIST	00:33	00:13	00:34	+1 s	+21 s
	DermaMNIST	01:12	02:12	01:15	+3 s	−57 s
	PathMNIST	11:52	28:46	12:26	+34 s	−980 s
	PneumoniaMNIST	00:57	02:03	01:02	+5 s	−61 s
	TissueMNIST	24:41	52:18	26:31	+110 s	−1547 s

Table 5: Epochs until early stopping and total training time (hh:mm:ss). **Bold**: shortest total time per row.

		SLPE		RoPE		SimPE	
Paradigm	Dataset	Ep.	Time	Ep.	Time	Ep.	Time
Fine-tuning	BloodMNIST	15	00:13:00	15	00:20:45	13	**00:11:55**
	BreastMNIST	18	00:00:54	35	00:01:10	10	**00:00:30**
	DermaMNIST	19	00:11:05	13	00:14:31	14	**00:08:10**
	PathMNIST	21	02:24:12	5	**00:37:10**	15	01:45:30
	PneumoniaMNIST	16	00:05:36	10	00:09:10	12	**00:04:24**
	TissueMNIST	21	04:30:57	48	08:28:48	15	**03:30:00**
From scratch	BloodMNIST	27	00:46:21	49	03:07:01	20	**00:35:20**
	BreastMNIST	25	00:13:45	40	**00:08:40**	22	00:12:28
	DermaMNIST	19	00:22:48	27	00:59:24	17	**00:21:15**
	PathMNIST	24	04:24:48	11	05:16:26	16	**03:22:16**
	PneumoniaMNIST	28	00:26:36	13	00:26:39	24	**00:24:48**
	TissueMNIST	24	09:52:24	52	45:19:36	22	**09:43:22**

Notably, on BreastMNIST scratch both SLPE and SimPE obtain a recall of 1.00 ± 0.00, indicating that both methods learn to predict the majority class when the positive class is too rare to learn from scratch. RoPE (0.56 ± 0.07) partially avoids this degenerate solution at the cost of lower precision.

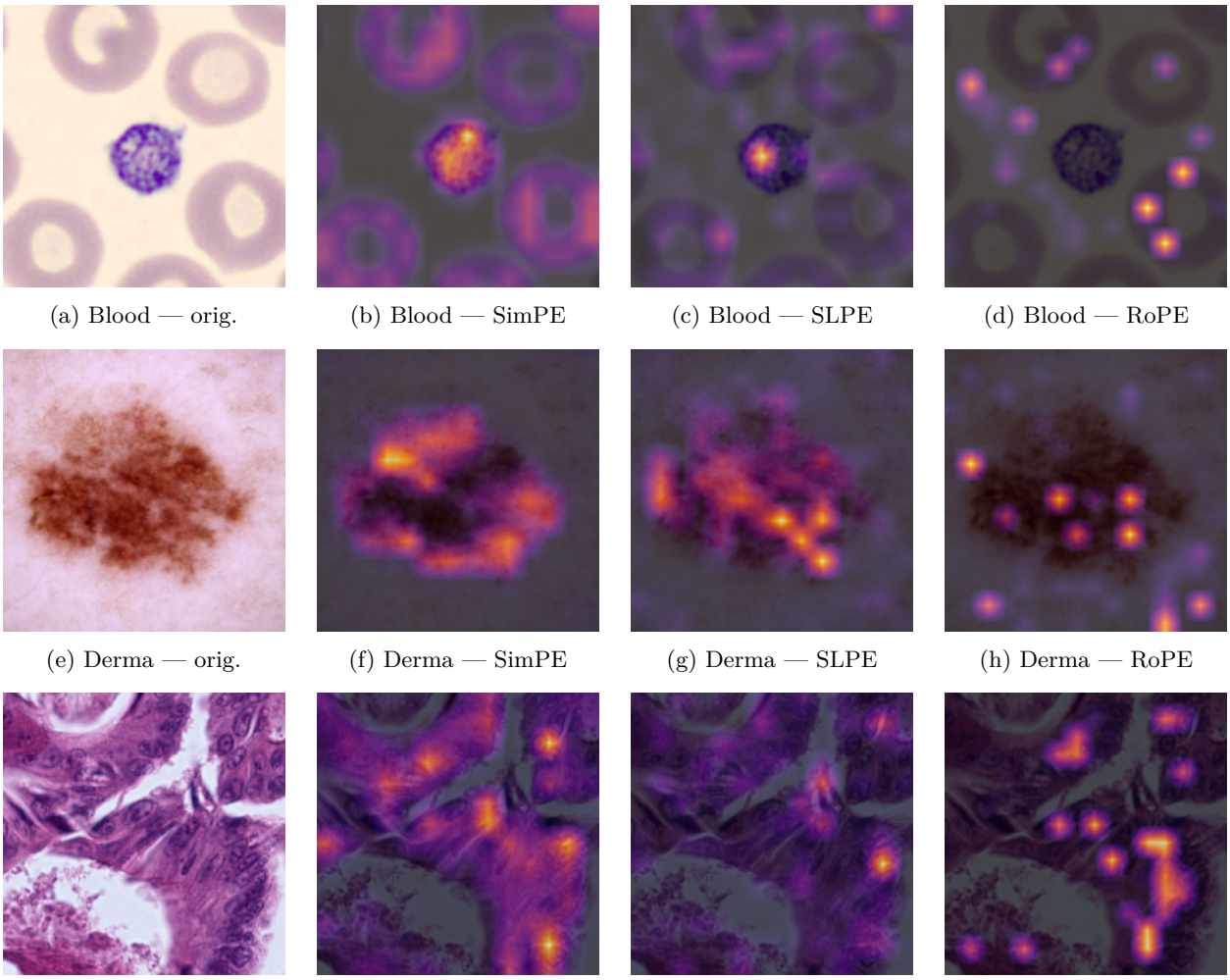
Overall. The advantage of SimPE is particularly evident in the scratch setting, where the absence of pre-trained representations makes the choice of positional encoding most critical. SLPE is consistently outperformed, while RoPE exhibits competitive but unstable behaviour across training regimes and datasets. Figure 6 in Appendix A visualises the full cross-validated distributions for all metrics.

5 Training Efficiency

5.1 Per-Epoch Training Time

The per-epoch overhead of SimPE over SLPE is minimal: 0–110 s across all 12 conditions, with 9 of 12 within 10 s. The exception is TissueMNIST, where the large dataset size amplifies per-sample overheads. The convolutional feature extraction in SimPE is lightweight and does not constitute a bottleneck.

SimPE's saving over RoPE is substantial in the from-scratch setting: PathMNIST 28 : 46 (RoPE) vs. 12 : 26 (SimPE) per epoch—a 2.3× speedup. This overhead arises because RoPE applies position-dependent rotations at every forward pass across all attention layers, while SimPE's extra computation is confined to the input stage.



(a) Blood — orig. (b) Blood — SimPE (c) Blood — SLPE (d) Blood — RoPE

(e) Derma — orig. (f) Derma — SimPE (g) Derma — SLPE (h) Derma — RoPE

(i) Path — orig. (j) Path — SimPE (k) Path — SLPE (l) Path — RoPE

Figure 1: Attention maps for BloodMNIST (top), DermaMNIST (middle), and PathMNIST (bottom). SimPE consistently focuses on clinically relevant regions: the platelet in Blood; the melanoma border in Derma; the dysplastic epithelium in Path. RoPE produces diffuse or misfocused activations.

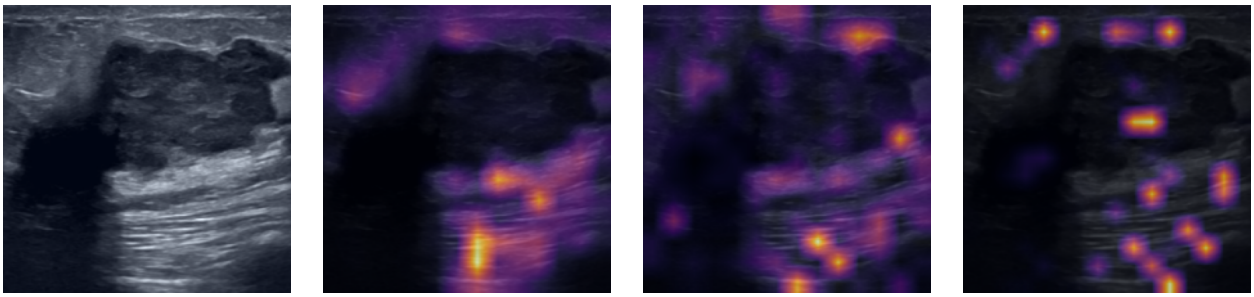
5.2 Convergence and Total Training Time

SimPE achieves the shortest total training time in 10 of 12 conditions. Despite its marginal per-epoch overhead over SLPE, it converges in 3–8 fewer epochs, yielding a net saving. The most extreme case is TissueMNIST from scratch: RoPE requires 45h 20m vs. 9h 52m for SLPE and 9h 43m for SimPE.

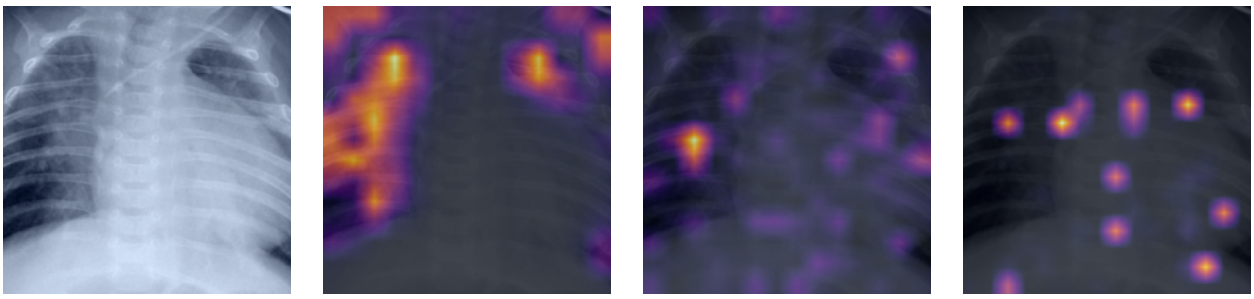
The two exceptions are PathMNIST fine-tuning (RoPE converges in 5 epochs but reaches only 90.14% accuracy vs. 98.14% for SimPE) and BreastMNIST from scratch (RoPE's very fast < 15 s per epoch lets it finish first despite 40 epochs). SimPE is therefore the most computationally efficient encoding when total wall-clock time is the criterion, while simultaneously delivering superior multi-metric performance.

6 Attention Maps Analysis

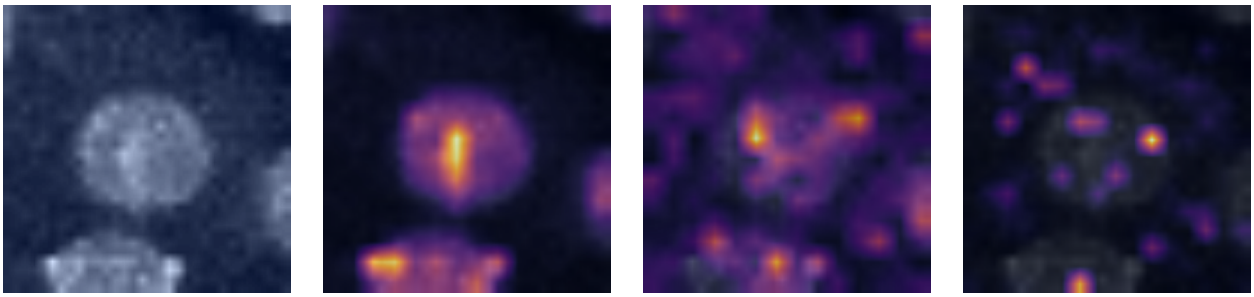
Figure 1 and Figure 2 show representative attention maps (original image, SimPE, SLPE, RoPE) across all six datasets.



(a) Breast — orig. (b) Breast — SimPE (c) Breast — SLPE (d) Breast — RoPE



(e) Pneumonia — orig. (f) Pneumonia — SimPE (g) Pneumonia — SLPE (h) Pneumonia — RoPE



(i) Tissue — orig. (j) Tissue — SimPE (k) Tissue — SLPE (l) Tissue — RoPE

Figure 2: Attention maps for BreastMNIST (top), PneumoniaMNIST (middle), and TissueMNIST (bottom). SimPE focuses on the lung fields in Pneumonia and the tubule structure in Tissue. All encodings are inadequate on BreastMNIST ultrasound, a known challenging dataset.

Across all six datasets, SimPE produces attention maps that are more spatially coherent and better aligned with clinically meaningful structures than either SLPE or RoPE. The following paragraphs provide a per-dataset analysis.

BloodMNIST. The image in Figure 1 (top row) shows a microscope slide containing a platelet—a small, circular, purplish structure at the centre of the field. SimPE's attention is tightly concentrated on that region, producing a coherent heatmap consistent with the classification label. SLPE yields a more diffuse activation spread across the entire image, with no clear focal point. RoPE almost completely ignores the platelet, activating elsewhere; this is consistent with its poor performance on this dataset in the from-scratch setting (Table 2).

DermaMNIST. The image shows a dermoscopic view of a skin melanoma. SimPE precisely attends to the irregular borders of the lesion—the very feature that dermatologists rely on for diagnosis (the ABCDE rule: Asymmetry, Border, Colour, Diameter, Evolution). SLPE focuses on the nevus body but without the specificity of SimPE. RoPE activates regions that do not even contain the lesion, making its attention map clinically uninterpretable.

PathMNIST. The image is a histopathological slide of colorectal adenocarcinoma epithelium. SimPE generates a broad, anatomically plausible focus on the dysplastic epithelial tissue, with homogeneous attention across the neoplastic region—consistent with the label "colorectal adenocarcinoma epithelium". SLPE produces a fragmented attention pattern with multiple sharp focal points but limited contextual integration. RoPE displays sparse and highly localised activations, with reduced sensitivity to the overall epithelial structure, which explains its severe performance degradation on PathMNIST in the scratch setting.

BreastMNIST. The ultrasound image shows a malignant breast tumour. All three attention maps are clinically inadequate: the models tend to focus on surrounding tissue rather than on the tumour region itself. Only RoPE shows a small activation area over the tumour, but at a level that is unsatisfactory for clinical use. This result is consistent with the overall lower performance on BreastMNIST and suggests that ultrasound images, with their characteristic speckle noise and low contrast-to-noise ratio, are challenging for all PE strategies irrespective of their design.

PneumoniaMNIST. The thoracic X-ray of a pneumonia patient shows SimPE clearly focusing on both lung fields, with only minor activations outside the pulmonary region. The attention remains consistent with the classification target. SLPE and RoPE both produce highly diffuse maps with no strongly activated region, offering little interpretive value.

TissueMNIST. The image depicts segments of a kidney proximal tubule under microscopy. SimPE precisely captures the tubular structures of interest. SLPE produces a highly diffuse attention with no clear anatomical target. RoPE generates an almost entirely absent activation, consistent with its poor performance on TissueMNIST (Table 2).

Summary. SimPE's geometry-aware encoding—based on pairwise patch similarity rather than absolute grid coordinates—guides the model towards features that are spatially structured and clinically meaningful. This interpretability advantage is independent of classification performance: even on BreastMNIST, where SimPE does not dominate numerically, its attention is no worse than the alternatives. On every other dataset it provides a qualitatively superior focus that supports the model's explainability.

7 Representation Quality via UMAP

Figures 3 and 4 show UMAP projections of the test-set `[CLS]` activations for all six datasets, grouped by the three positional encodings.

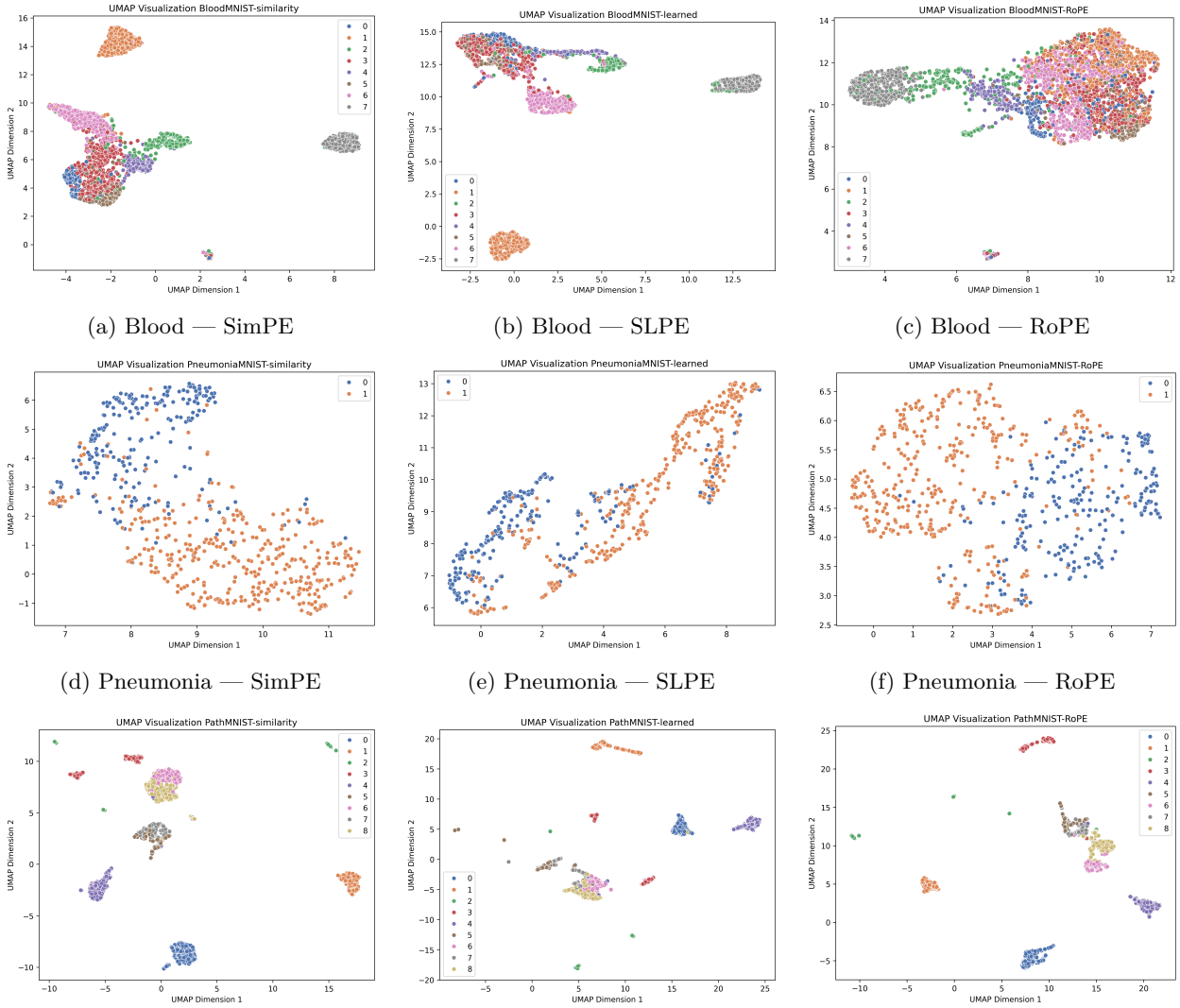
BloodMNIST. SimPE separates morphologically distinct cell types clearly, notably neutrophils and lymphocytes. SLPE produces overlapping manifolds. RoPE collapses most classes into a single cluster, blending classes 1 and 7 that are otherwise well separated.

PneumoniaMNIST. SimPE produces two compact, well-separated lobes for the binary classes. SLPE and RoPE show broader manifolds with larger overlap, directly explaining the higher recall of SimPE on this dataset (Table 3).

PathMNIST. SimPE reveals biologically coherent structure: mucosa and smooth muscle (classes 5/7) and adipose with background (classes 6/8) form adjacent but distinct clusters, and the normal colon mucosa class (class 0) is completely isolated. Neither SLPE nor RoPE reproduces this grouping.

DermaMNIST. SLPE splits the most prevalent class (vascular lesions, class 5) into two disconnected groups, indicating that it conflates lesions with different textures. SimPE and RoPE avoid this artefact.

BreastMNIST. SimPE shows slightly better grouping for the normal class (class 0), but differences are modest and consistent with the lower overall performance on this challenging dataset.



(a) Blood — SimPE (b) Blood — SLPE (c) Blood — RoPE

(d) Pneumonia — SimPE (e) Pneumonia — SLPE (f) Pneumonia — RoPE

(g) Path — SimPE (h) Path — SLPE (i) Path — RoPE

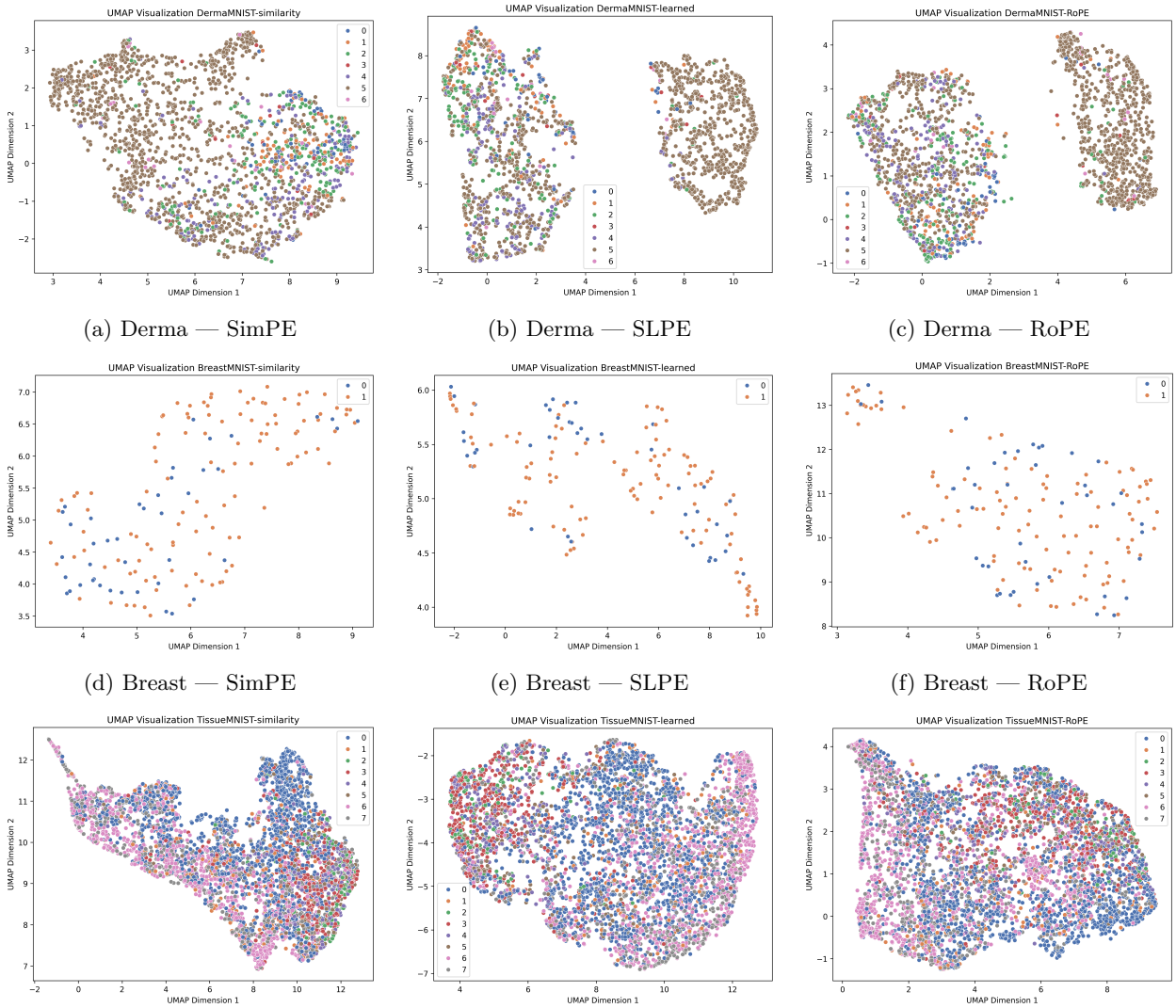
Figure 3: UMAP projections for BloodMNIST (top), PneumoniaMNIST (middle), and PathMNIST (bottom). SimPE (left) consistently yields tighter, better-separated clusters compared to SLPE (centre) and RoPE (right).

TissueMNIST. No meaningful differences across encodings, consistent with the dataset's large size (236,386 instances) and high within-class variance—which dominate the learned representations regardless of PE choice.

Summary. SimPE encodes *geometric discriminative structure* into the feature representations beyond decision boundaries. Biologically meaningful cluster organisation supports downstream tasks such as retrieval, few-shot transfer, and active learning—capabilities increasingly important in low-resource clinical settings.

8 Extension to 3D Medical Images

Current approaches for 3D medical image classification with ViTs use either CNN preprocessing (Jia & Shu, 2021) (high parameter count) or multi-slice aggregation (Jang & Hwang, 2022) (quadratic attention growth).



(a) Derma — SimPE (b) Derma — SLPE (c) Derma — RoPE

(d) Breast — SimPE (e) Breast — SLPE (f) Breast — RoPE

(g) Tissue — SimPE (h) Tissue — SLPE (i) Tissue — RoPE

Figure 4: UMAP projections for DermaMNIST (top), BreastMNIST (middle), and TissueMNIST (bottom).

SimPE avoids both issues by simply replacing 2D with 3D convolutions in the feature extraction step—no other change to the ViT is required.

SimPE is compared against SLPE on four 3D MedMNIST datasets (Yang et al., 2023): **NoduleMNIST3D** (Armato et al., 2011) (CT pulmonary nodules, from LUNA16 (Setio et al., 2017)), **AdrenalMNIST3D** (CT adrenal glands, from KiTS (Heller et al., 2021)), **OrganMNIST3D** (abdominal CT, from TCIA (Roth et al., 2015)), and **FractureMNIST3D** (Lin et al., 2023). RoPE was excluded because the `timm` library (Wightman, 2019) provides no 3D Transformer with RoPE support; a custom implementation would undermine consistency.

On NoduleMNIST3D and AdrenalMNIST3D, SLPE exhibits a degenerate failure: zero recall and F1, predicting only the majority class throughout.[2] SimPE avoids this collapse on all four datasets. On OrganMNIST3D the margin is dramatic (64.8% vs. 21.5% accuracy). These results confirm that SimPE's geometry-aware inductive bias benefits 3D medical image classification with minimal additional implementation effort.

[2]Precision is conventionally set to 1.0 for this degenerate case.

Table 6: SimPE vs. SLPE on 3D MedMNIST (benchmark split, 95% CI). Acc. = Accuracy (%), P/R/F1 are scores.

	SimPE				SLPE			
Dataset	**Acc.**	**P**	**R**	**F1**	**Acc.**	**P**	**R**	**F1**
NoduleMNIST3D	88.1 ± 1.0	0.73	0.69	0.66	79.4 ± 1.0	1.00	0.00	0.00
AdrenalMNIST3D	81.5 ± 1.2	0.64	0.34	0.44	76.8 ± 1.3	1.00	0.00	0.00
OrganMNIST3D	64.8 ± 1.4	0.67	0.66	0.63	21.5 ± 1.1	0.11	0.17	0.12
FractureMNIST3D	55.4 ± 1.3	0.41	0.49	0.44	53.8 ± 1.2	0.39	0.47	0.42

9 Conclusion

This paper has provided a comprehensive evaluation of SimPE across dimensions previously unaddressed: multi-metric cross-validated performance, per-epoch and total training cost, attention map interpretability, representation quality, and extension to 3D volumes.

SimPE outperforms SLPE and RoPE on most precision, recall, and F1 metrics in addition to accuracy, with narrower confidence intervals indicating more stable training. Its per-epoch cost is marginally higher than SLPE but substantially lower than RoPE, and faster convergence makes it the most efficient choice overall. Attention maps show that SimPE focuses on clinically relevant structures more reliably. UMAP projections reveal tighter, biologically coherent class clusters. SimPE extends seamlessly to 3D volumetric images by replacing 2D with 3D convolutions, with strong improvements over SLPE.

Taken together, these findings establish SimPE as a practically superior positional encoding strategy for geometry-structured medical image classification, combining accuracy, stability, efficiency, interpretability, and representation quality.

Future work will investigate whether the representation quality advantage translates into improved few-shot generalisation under domain shift—a clinically critical scenario for low-resource medical institutions. The source code repository will be made publicly available upon acceptance.

Broader Impact Statement

SimPE improves accuracy, interpretability, and computational efficiency of ViTs for medical image classification, with potential to support diagnostic workflows in radiology, pathology, and dermatology. We caution that all results use MedMNIST benchmarks and may not generalise to all clinical populations or imaging equipment; further validation on institution-specific data is required before deployment. SimPE should be treated as a decision-support tool, not a replacement for clinical expertise.

References

Andrea Acevedo and Anna Merino. A dataset of microscopic peripheral blood cell images for development of automatic recognition systems. *Data in Brief*, 30:105474, 2020.

Walid Al-Dhabyani, Mohammed Gomaa, and et al. Dataset of breast ultrasound images. *Data in Brief*, 28: 104863, 2020.

Samuel G. Armato, Geoffrey McLennan, and et al. The lung image database consortium (LIDC) and image database resource initiative (IDRI): A completed reference databaseof lung nodules on CT scans. *Medical Physics*, 38(2):915–931, 2011.

B.E. Bejnordi, M. Veta, P.J. van Diest, B. van Ginneken, N. Karssemeijer, G. Litjens, and J.A.W.M. van der Laak. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Journal of the American Medical Association*, 318:2199–2210, 2017.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255. IEEE, 2009. doi: 10.1109/CVPR.2009.5206848.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021. URL `https://openreview.net/forum?id=YicbFdNTTy`.

Nicholas Heller, Fabian Isensee, Klaus H Maier-Hein, Xiangde Hou, Cong Xie, Feng Li, Yunlu Nan, Guiqiu Mu, Ziyi Lin, Man Han, et al. The state of the art in kidney and kidney tumor segmentation in contrast-enhanced ct imaging: Results of the kits19 challenge. *Medical Image Analysis*, 67:101821, 2021. doi: 10.1016/j.media.2020.101821.

Habib Irani and Vangelis Metsis. Positional encoding in transformer-based time series models: A survey, 2025. URL `https://arxiv.org/abs/2502.12370`.

Jinseong Jang and Dosik Hwang. M3t: three-dimensional medical image classifier using multi-plane and multi-slice transformer. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20686–20697, 2022. doi: 10.1109/CVPR52688.2022.02006.

Qiang Jia and Heng Shu. Bitr-unet: A cnn-transformer combined network for mri brain tumor segmentation. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pp. 3–14. Springer, September 2021. doi: 10.1007/978-3-031-09002-8_1. Epub 2022 Jul 15.

Kai Jiang, Peng Peng, Youzao Lian, and Weisheng Xu. The encoding method of position embeddings in vision transformer. *Journal Visual Comunication and Image Representation*, 89(C), November 2022. ISSN 1047-3203. doi: 10.1016/j.jvcir.2022.103664. URL `https://doi.org/10.1016/j.jvcir.2022.103664`.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

G. Leonardi, L. Portinale, and A. Santomauro. Similarity-based positional encoding for enhanced classification in medical images. In *Proc. 3rd AIxIA Workshop on Artificial Intelligence For Healthcare*, pp. 182–188, 2024. URL `https://ceur-ws.org/Vol-3880/paper16.pdf`. CEUR Workshop Proceedings.

Yuheng Li, Mingzhe Hu, Richard L. J. Qiu, Maria Thor, Andre Williams, Deborah Marshall, and Xiaofeng Yang. Romedformer: A rotary-embedding transformer foundation model for 3d genito-pelvic structure segmentation in MRI and CT, 2025. URL `https://arxiv.org/abs/2503.14304`.

Hui Ming Lin, Errol Colak, Tyler Richards, Felipe C Kitamura, Luciano M Prevedello, Jason Talbott, Robyn L Ball, Ekim Gumeler, Kristen W Yeom, Mazin Hamghalam, et al. The rsna cervical spine fracture ct dataset. *Radiology: Artificial Intelligence*, 5(5):e230034, 2023. doi: 10.1148/ryai.230034.

Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations (ICLR)*, 2017.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *Journal of Open Source Software*, 3(29):861, 2018. doi: 10.21105/joss.00861. URL `https://doi.org/10.21105/joss.00861`.

Holger R Roth, Le Lu, Ali Farag, Ho-Cheol Shin, Jiamin Liu, Evrim B Turkbey, and Ronald M Summers. Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 556–564. Springer, 2015. doi: 10.1007/978-3-319-24553-9_68.

Andrea Santomauro, Giorgio Leonardi, and Luigi Portinale. Comparing different positional encodings for the interpretation of medical images. In *Communications in Computer and Information Science*, volume 2830, pp. 78–89. Springer, 2026. ISBN 978-3-032-16708-8.

Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Geert Litjens, Sarah Gerke, Colin Jacobs, Sander J van Riel, Mathilde M Wille, Mahdi Naqibullah, and Bram van Ginneken. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The luna16 challenge. *Medical Image Analysis*, 42:1–13, 2017. doi: 10.1016/j.media.2017.06.015.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding, 2023. URL `https://arxiv.org/abs/2104.09864`.

Yucheng Tang, Dong Yang, Wenqi Li, Holger R. Roth, Bennett Landman, Daguang Xu, Vishwesh Nath, and Ali Hatamizadeh. Self-supervised pre-training of swin transformers for 3d medical image analysis. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20698–20708, 2022. doi: 10.1109/CVPR52688.2022.02007.

Philipp Tschandl. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions, 2018. URL `https://doi.org/10.7910/DVN/DBW86T`.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *31st Conference on Neural Information Processing Systems (NIPS)*, Long Beach, CA, 2017. URL `https://tinyurl.com/5463rwkz`.

Ljosa Vebjorn, Katherine L Sokolnicki, and et al. Annotated high-throughput microscopy imagesets for validation. *Nature methods*, 9(7):637–637, 2012.

Libo Wang, Rui Li, Ce Zhang, Shenghui Fang, Chenxi Duan, Xiaoliang Meng, and Peter M. Atkinson. Unetformer: A unet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 190:196–214, 2022. ISSN 0924-2716. doi: https://doi.org/10.1016/j.isprsjprs.2022.06.008. URL `https://www.sciencedirect.com/science/article/pii/S0924271622001654`.

Luyao Wang, Xiaoyan Wang, Bangze Zhang, Xiaojie Huang, Cong Bai, Ming Xia, and Peiliang Sun. Multi-scale hierarchical transformer structure for 3d medical image segmentation. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 1542–1545, 2021. doi: 10.1109/BIBM52615.2021.9669799.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. ChestX-ray14, 2017. URL `https://paperswithcode.com/dataset/chestx-ray14`.

Ross Wightman. Pytorch image models. `https://github.com/huggingface/pytorch-image-models`, 2019. Accessed: 2025-10-17.

Kan Wu, Houwen Peng, Minghao Chen, Jianlong Fu, and Hongyang Chao. Rethinking and improving relative position encoding for vision transformer. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10013–10021, 2021. doi: 10.1109/ICCV48922.2021.00988.

Jiancheng Yang, Rui Shi, and Bingbing Ni. Medmnist classification decathlon: A lightweight automl benchmark for medical image analysis. In *IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pp. 191–195, 2021.

Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41, 2023.

Zhengyang Yu and Jochen Triesch. CRE: Circle relationship embedding of patches in vision transformer. *ESANN 2023 proceesdings*, 2023. doi: 10.14428/esann/2023.es2023-75.

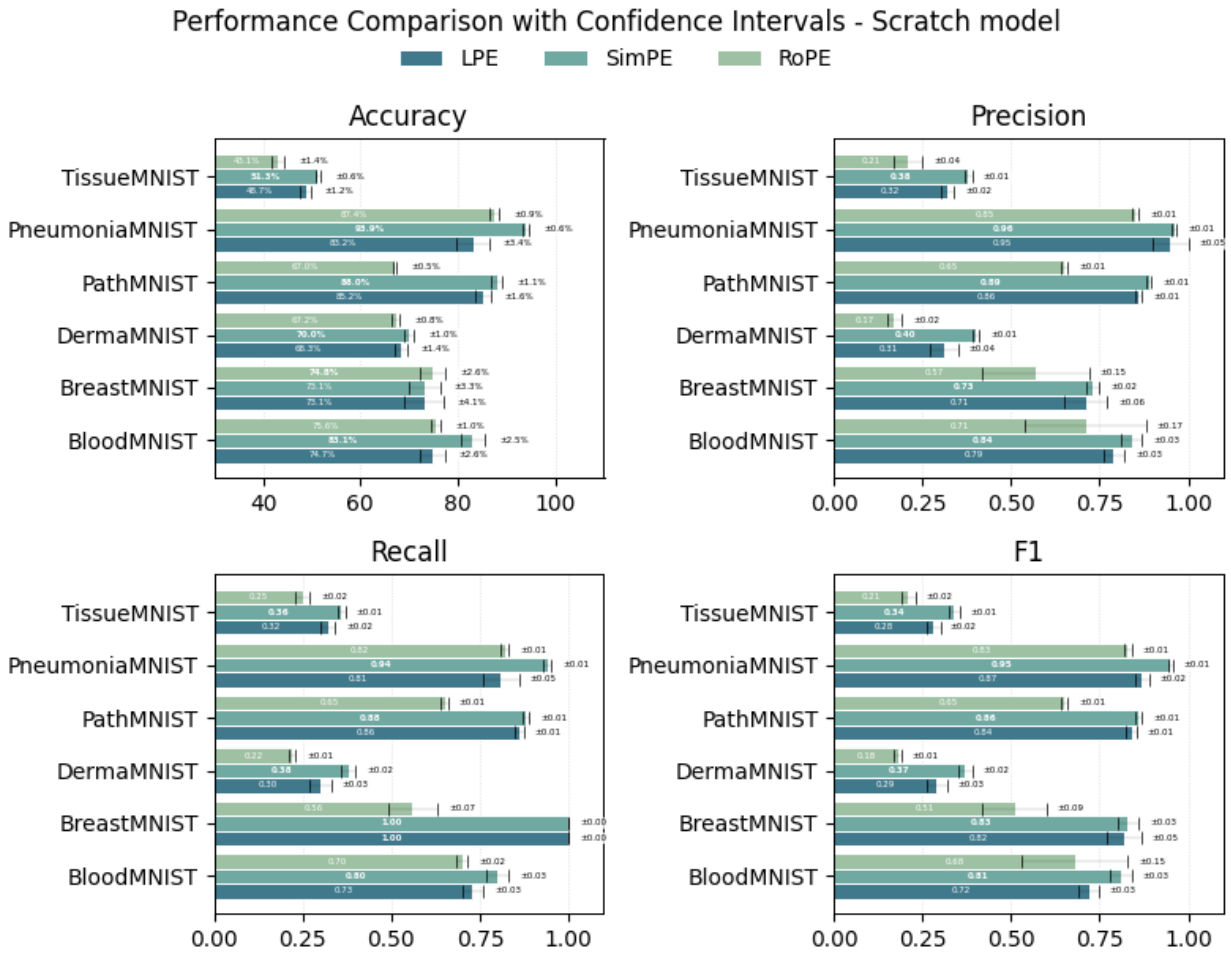


Figure 5: Mean ± 95% CI from 10-fold CV for the scratch-trained model (SLPE, RoPE, SimPE across accuracy, precision, recall, F1).

A Cross-Validation Performance Charts

B Reproducibility

All experiments used a single NVIDIA A100 GPU. The complete hyperparameter configuration is given in Table 1. Source code, model configurations, and training scripts will be made publicly available upon acceptance of this paper (link omitted for double-blind review).

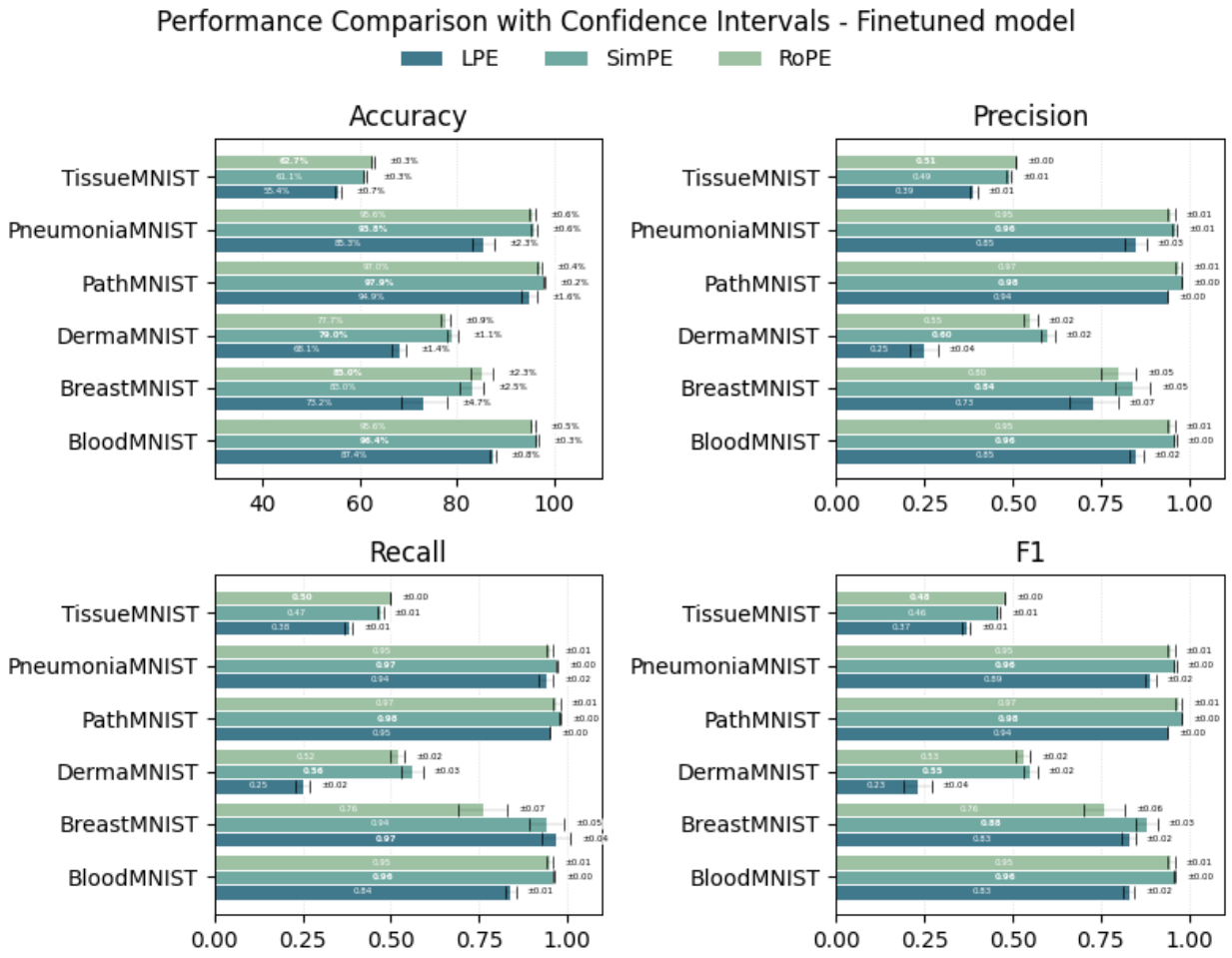


Figure 6: Mean ± 95% CI from 10-fold CV for the fine-tuned model.