Spatially refined transformer embeddings for accurate histopathology tissue classification

Przemysław Niedziela[1][0009−0005−3668−8446] and Bogusław Cyganek[1][0000−0001−5185−1145]

AGH University of Krakow, Al. Mickiewicza 30, 30-059 Kraków, Poland
{niedziel,cyganek}@agh.edu.pl

Abstract. Accurate recognition of tissue types in histopathological whole-slide images is a fundamental challenge in computational pathology, particularly given the global decline in practicing histopathologists. We present a classification framework that integrates a lightweight Vision Transformer (TinyViT) for patch-level feature extraction with dimensionality reduction via Principal Component Analysis and Neighborhood Component Analysis. The resulting low-dimensional yet discriminative representations are classified using k-Nearest Neighbors and Support Vector Machines, yielding state-of-the-art performance on the DiagSet prostate cancer dataset and on BreakHis, even under extreme reductions in feature dimensionality. To further improve robustness, we introduce a spatial refinement strategy that projects patch predictions into a grid representation of the slide, enforcing spatial consistency by identifying and reclassifying low- and high-confidence regions. This two-stage process enhances predictive accuracy and improves interpretability by highlighting confident as well as uncertain tissue areas. On DiagSet, our framework achieves 82.70% in the 4-class and over 90.6% in the binary setting, surpassing prior baselines, while post-hoc spatial refinement yields consistent gains of up to 2 percentage points without retraining.

Keywords: Computational Pathology · Vision Transformers, TinyViT · Neighborhood Component Analysis (NCA) · Whole-Slide Imaging (WSI) · Dimensionality Reduction · Spatial Refinement · Prostate Cancer.

1 Introduction

The increasing disparity between the rapidly growing number of patients and the declining availability of histopathologists poses a significant challenge to modern healthcare. Artificial intelligence offers a promising solution by enabling automated, accurate, and scalable medical image analysis. However, methods must deliver high accuracy, computational efficiency, and robustness across heterogeneous data sources. Earlier approaches to image classification relied on handcrafted descriptors such as SIFT combined with bag-of-words models [19]. More recently, transformer-based vision models have advanced medical image analysis [3, 16, 17], with self-supervised variants (e.g., DINO) [8] showing strong

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_31

feature extraction capability. Notable applications include Surgical-DINO [11] and the Medical Slice Transformer [12], which often match or surpass CNNs [7, 9, 5, 6]. Yet training large self-supervised models can be computationally demanding. TinyViT [2] offers a compact alternative that preserves much of the accuracy while improving efficiency.

This work makes two complementary contributions. First, we develop a lightweight classification framework for histopathology that couples TinyViT-5M with a tailored feature pipeline: following the aggregation schema of ResFeats [18], we concatenate *multi-stage* TinyViT features and apply dimensionality reduction via PCA and Neighborhood Component Analysis (NCA) [13], followed by kNN/SVM classifiers. Second, we introduce a *post-hoc spatial refinement* strategy that leverages coordinates *only at test time*. Patch-level probabilities are transformed into signed log-odds and aggregated on a stride-aligned grid using a difference-array (summed-area) scheme [15] to produce an *aggregated evidence field* $E_{\text{agg}}(x, y)$. Within this field, we identify low- and high-evidence regions via thresholding and re-predict patches in those neighborhoods, enforcing spatial coherence without retraining.

We evaluate on two benchmarks - the DiagSet prostate cancer dataset [1] and BreakHis [14]. Using pre-extracted patches, our base pipeline attains state-of-the-art-level performance. On re-extracted, spatially indexed DiagSet patches, the same trained model benefits further from refinement, improving accuracy while adding limited overhead.

Together, these findings demonstrate that reliable tissue classification can be performed within a substantially reduced feature space while leveraging spatial context for error correction. This opens possibilities for processing very large datasets [5, 6] and supports the development of lightweight AI models in medical imaging [4]. Moreover, by modeling spatial consistency and highlighting confidence regions, our approach provides structured information readily integrable into large language models.

Our main contributions are:

1. **Efficient architecture:** TinyViT-5M with multi-stage feature concatenation and supervised dimensionality reduction (PCA+NCA) for compact, discriminative representations.
2. **Post-hoc spatial refinement:** A coordinate-aware, test-time procedure that aggregates signed log-odds via a difference-array and triggers confidence-guided re-prediction.
3. **Strong empirical performance:** kNN/SVM back-ends achieve state-of-the-art-level results on DiagSet and BreakHis.
4. **Extremely short features:** High accuracy retained with very few components, supporting resource-constrained deployment.
5. **Interpretability:** The aggregated evidence field provides an interpretable spatial map of model confidence, enabling visualization of uncertain regions.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_31

2 Materials and methods

2.1 Datasets

DiagSet [1] is a large-scale prostate cancer histopathology WSI collection comprising 5151 slides in three subsets (A, B, C). DiagSet-A includes over 2.6 million patches from 430 fully annotated slides (Fig. 1), extracted at four magnification levels (5×, 10×, 20×, 40×) using 256×256 px windows with stride 128. Each patch carries one of nine labels: scan background, tissue background, normal tissue, acquisition artifact, or Gleason grades R1–R5. Our experiments focus on subset A.1 at 5× under two settings: a *4-class* setting where the non-cancerous class (NC) combines A, N, T, R1 and R2, with remaining classes representing R3, R4 and R5; and a *binary* setting grouping R1–R5 as cancerous vs. NC. We introduce mild class balancing via majority-class downsampling (Fig. 2), with scans retained within their respective splits. The original test split is used for SOTA comparison.

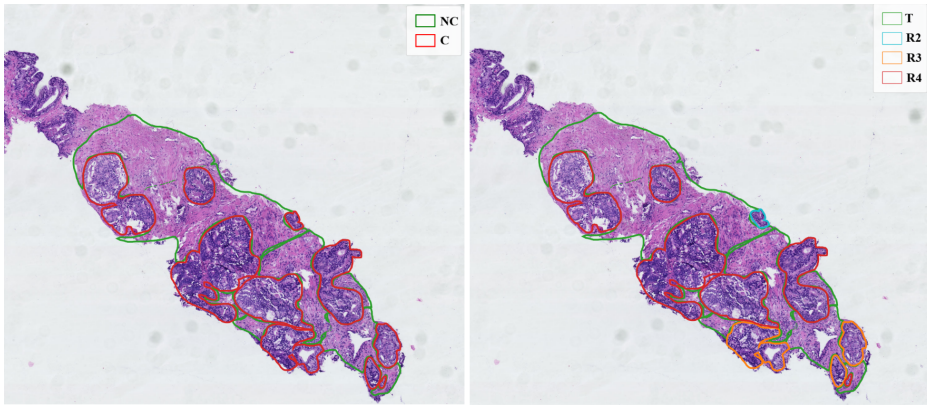

Fig. 1: Selected WSI region with binary annotations (left) and 4-class annotations (right).

BreakHis [14] is a publicly available breast cancer dataset containing 7,909 H&E-stained microscopic images from 82 patients at four magnifications (40×, 100×, 200×, 400×), all 700×460 px. Each image is annotated as benign or malignant with further subdivision into eight subtypes (four per category). We employ the multiclass task with stratified 5-fold cross-validation and majority-class undersampling.

2.2 Proposed method

Backbone and pretraining. TinyViT [2] is a compact Vision Transformer composed of four sequential stages (Fig. 3), each progressively reducing spatial resolution. Patch embedding uses a 2D convolution (kernel 3, stride 2, padding 1). Stage 1 employs convolutions for low-level features; Stages 2–4 use transformer blocks with window-based self-attention. We use TinyViT-5M (∼5M parameters; depths 2, 2, 6, 2; embedding dims 64, 128, 160, 320).

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_31

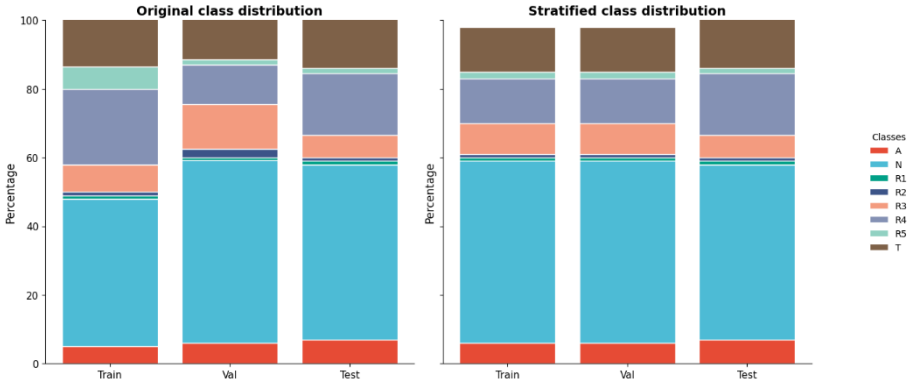

Fig. 2: Class distribution in original (left) and downsampled DiagSet (right).

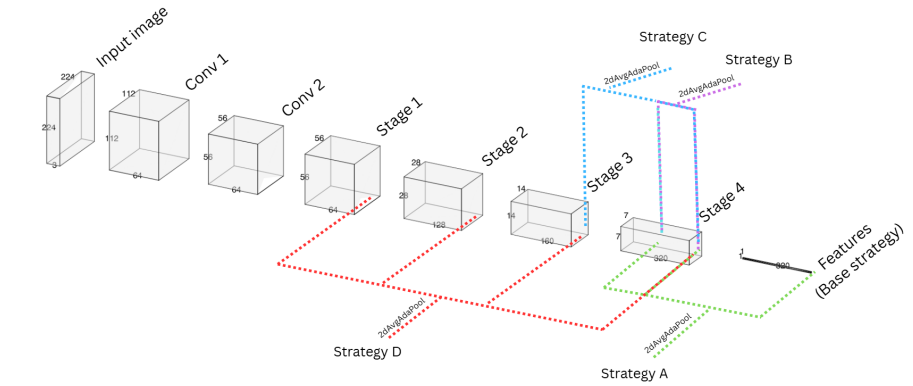

Fig. 3: Feature flow in TinyViT with sampling points for each extraction strategy.

TinyViT-5M is pretrained on ImageNet-22k and fine-tuned on ImageNet-1k. We then fine-tune on *DiagSet-A.1* for 50 epochs using a cosine learning-rate schedule with 5 warm-up epochs (base LR 0.1, warm-up LR 2×10^{-7}). A WeightedRandomSampler compensates for class imbalance, and standard augmentations are applied: flips (p=0.5), resized crops to 224×224, rotations (±30°), and random erasing (p=0.5).

For BreakHis, we fine-tune TinyViT-5M *independently per magnification* using 5-fold cross-validation: 100 epochs, 15 warm-up epochs (warm-up LR 2×10^{-5}), same augmentations and class-imbalance handling.

After fine-tuning, we extract features from multiple hierarchical levels. All intermediate maps are pooled to 1×1 via adaptive average pooling before concatenation. We define five strategies (sampling points in Fig. 3):

1. **Base** (*320-dim*): pooled feature from the classification head.
2. **A** (*960-dim*): concatenation of the pooled head and both Transformer blocks in Stage 4.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_31

3. **B** (*640-dim*): concatenation of the two pooled Stage 4 blocks.
4. **C** (*800-dim*): concatenation of pooled Stage 3 output and both Stage 4 blocks.
5. **D** (*672-dim*): concatenation of pooled outputs from all four stages.

Multi-stage feature concatenation was demonstrated for CNNs in [18]; this paper applies the approach to vision transformers for the first time, achieving even higher accuracies than classical ViT, which constitutes one of our main contributions.

Dimensionality reduction and feature compressions Dimensionality reduction constitutes a critical step in the experimental pipeline, as it enables both computational efficiency and the extraction of more compact and highly discriminative representations. In this work, we employ two complementary methods: Principal Component Analysis (PCA) and Neighborhood Component Analysis (NCA) [13]. PCA is an unsupervised technique that identifies the directions in the data along which variance is maximized. By projecting features onto these directions, PCA reduces redundancy and noise while preserving the most informative global structure of the data. In contrast, NCA is a supervised method that explicitly seeks to improve classification performance. Rather than focusing on variance, NCA learns a feature transformation that maximizes the likelihood that samples from the same class are close to one another in the reduced space. Within our framework, PCA and NCA jointly provide an effective strategy for compressing high-dimensional TinyViT embeddings into compact and computationally manageable representations. To assess the efficacy of these methods, we systematically applied feature compression across multiple stages of the pipeline, varying the number of retained components from the full embedding dimensionality down to a single component. This evaluation allowed us to quantify how well discriminative information is preserved under progressive dimensionality reduction.

Classifiers Following feature extraction and dimensionality reduction, classification is performed using two complementary methods: k-Nearest Neighbors (kNN) and Support Vector Machines (SVM). KNN is a non-parametric method that assigns a class label to a new sample based on the labels of its closest neighbors in the feature space. This approach directly exploits the local structure of the data, making it particularly effective when combined with NCA, which explicitly optimizes the embedding space for neighborhood-based classification. In our context, kNN serves as a straightforward yet powerful tool to evaluate how well the compressed features cluster according to tissue type.

SVM, by contrast, constructs decision boundaries that maximize the margin between classes. Unlike kNN, which relies on local proximity relationships, SVM identifies global separating hyperplanes that generalize well even in the presence of partial class overlap. Within our framework, the SVM classifier complements kNN by providing a margin-based decision strategy that benefits from

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_31

both PCA's variance-preserving projection and NCA's discriminative subspace optimization. Together, these classifiers provide a robust evaluation of the representational quality of the compressed transformer embeddings.

Pipeline architecture. Figure 4 summarizes the framework: (1) dataset preparation and model training, (2) feature extraction with TinyViT-5M, (3) compression via PCA/NCA, and (4) classification with kNN/SVM.

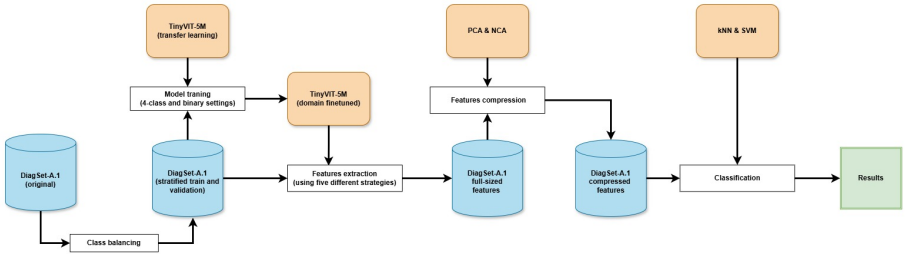

Fig. 4: Experimental pipeline: from raw data through feature extraction and compression to classification.

2.3 Post-hoc spatial refinement

The refinement procedure was conducted exclusively for the binary (2-class) configuration of the DiagSet dataset. This design choice reflects the mathematical formulation of the refinement mechanism, which aggregates signed log-odds evidence between two opposing classes. Extending this framework to multi-class settings would require a separate one-vs-all or pairwise decomposition of class probabilities, substantially increasing computational complexity and reducing the interpretability of the resulting evidence maps. Restricting the analysis to the binary case thus ensures a consistent probabilistic interpretation of a local evidence.

Following initial classification, a spatial refinement stage was introduced to improve local consistency and reduce patch-level noise in WSI predictions. For this step, we re-extracted patches from the same test WSIs using the official dataset code and publicly available tissue annotations, reproducing the original sampling protocol and ensuring spatial correspondence with the benchmark data. Although exact pixel-level alignment with the official pre-extracted patches cannot be guaranteed, the resulting set maintains an equivalent coverage and geometry of tissue regions. In contrast to the training data, which contained only pre-extracted patches without coordinates, these newly generated patches preserve their spatial metadata. The trained TinyViT-based feature extractor and classifiers remain unchanged; coordinates are used exclusively at test time to perform the refinement procedure described below.

We propose a spatial aggregation framework to quantify and visualize *local model reliability* (how consistently and confidently a model's predictions agree

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_31

within a small spatial neighborhood) from patch-level classification results. The method operates on sets of patch predictions (x_i, y_i, p_i, l_i), where (x_i, y_i) are the spatial coordinates of patch i, $p_i \in (0, 1)$ is the predicted probability for the positive class, and $l_i \in \{0, 1\}$ is the corresponding predicted label. The goal is to transform these discrete patch predictions into a continuous *evidence field* $E_{\text{agg}}(x, y)$, which represents local consensus and confidence across overlapping patches. Regions are subsequently identified based on evidence and flagged for re-prediction or review.

The first step converts patch-level probabilities p_i into *log-odds values* (logits) - continuous, linear magnitudes of evidence uncertainties:

$$L_i = \log \frac{p_i}{1 - p_i} \tag{1}$$

To prevent numerical divergence, probabilities are clipped to $p_i \leftarrow \min(\max(p_i, \varepsilon), 1-\varepsilon)$ with $\varepsilon = 10^{-4}$. This transformation yields $L_i > 0$ for evidence favoring the positive class and $L_i < 0$ otherwise, while $|L_i|$ encodes the magnitude of model confidence. To represent both classes in a unified evidence domain, the binary label $l_i \in \{0, 1\}$ is mapped to a signed indicator variable $s_i \in \{-1, +1\}$:

$$s_i = 2l_i - 1. \tag{2}$$

Each patch's signed evidence is then defined as

$$E_i = s_i \cdot L_i. \tag{3}$$

This ensures that both classes contribute to the evidence map with opposing signs, where large positive E_i denotes strong positive-class support and large negative E_i strong negative-class support.

Given patch coordinates (x_i, y_i), we define a discrete spatial grid with stride S and patch size P. Each patch contributes its evidence value E_i to a local footprint of size

$$s = \left\lceil \frac{P}{S} \right\rceil, \tag{4}$$

spanning $s \times s$ neighboring grid cells.

To efficiently aggregate overlapping patches, we employ a *summed-area table* technique [15], which computes local sums in $O(N + HW)$ time, where N is the number of patches and $H \times W$ the grid size. Let D_{sum} and D_{wgt} denote zero-initialized arrays representing difference accumulators for total evidence and patch counts, respectively. For each patch located at grid coordinates (i_0, j_0), we apply:

$$\begin{aligned} D_{\text{sum}}[j_0, i_0] &+= E_i, \\ D_{\text{sum}}[j_0 + s, i_0] &-= E_i, \\ D_{\text{sum}}[j_0, i_0 + s] &-= E_i, \\ D_{\text{sum}}[j_0 + s, i_0 + s] &+= E_i, \end{aligned} \tag{5}$$

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_31

and similarly for D_{wgt} with unit increments.

The cumulative evidence and weights are obtained by 2D integration:

$$\text{SumGrid} = \text{cumsum}_x(\text{cumsum}_y(D_{\text{sum}})), \quad \text{WgtGrid} = \text{cumsum}_x(\text{cumsum}_y(D_{\text{wgt}})). \tag{6}$$

The resulting mean class-signed evidence field is then computed as

$$E_{\text{agg}}(x, y) = \frac{\text{SumGrid}(x, y)}{\max(\text{WgtGrid}(x, y), \varepsilon)}. \tag{7}$$

This operation produces a smooth field capturing local consensus and disagreement between overlapping patches. The evidence field $E_{\text{agg}}(x, y)$ is analyzed to identify regions of low and high local confidence. A binary mask is first defined by thresholding in log-odds space:

$$M(x, y) = \begin{cases} 1, & E_{\text{agg}}(x, y) > \tau, \\ 0, & \text{otherwise}, \end{cases} \tag{8}$$

where τ is a chosen threshold (in our work fixed to $\tau = 0.99$). Since $E_{\text{agg}}(x, y)$ represents aggregated log-evidence rather than direct probability, the threshold $\tau = 0.99$ corresponds to moderately strong positive evidence. This value was empirically selected to isolate spatial regions that exhibit consistent positive consensus across overlapping patches, while excluding weak or conflicting evidence.

Following thresholding, each patch anchor (x_k, y_k) is evaluated within a local 2×2 neighborhood. If any location within this neighborhood satisfies $M(x, y) = 1$, the corresponding patch is marked for (positive) re-prediction, otherwise set to 0:

$$r_k = \begin{cases} 1, & \displaystyle\sum_{(x,y) \in \mathcal{N}(x_k, y_k)} M(x, y) > 0, \\ 0, & \text{otherwise}. \end{cases} \tag{9}$$

This operation enforces local spatial coherence by re-evaluating patches adjacent to confident regions in the binary mask $M(x, y)$.

To isolate the effect of refinement, we conduct a pure ablation: refinement operates only at post-processing time on the unchanged predictions and probabilities produced by the baseline pipeline. Model weights, inputs, and inference configuration are identical to the non-refined setting.

Beyond improving spatial coherence, the resulting confidence and uncertainty maps offer a structured contextual representation suitable for downstream integration with large language models, enabling future multimodal approaches that combine visual evidence with textual reasoning.

3 Results

All experiments primarily involve the DiagSet dataset. BreakHis is used solely for cross-dataset verification.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_31

3.1 Classification without spatial context

We evaluate TinyViT features under five extraction strategies (Base, A–D) and report results organized by projection method and classifier. For space efficiency, Tables 1–2 present Base and Strategy A (the most consistently competitive strategy) across all pipelines, with best Strategy B–D results noted in footnotes.

4-class setting. The strongest 4-class configuration is NCA+SVM with Strategy D, reaching 82.70% at 30 components. Competitive results are observed for NCA+kNN (Base) at 82.63% (size 80) and PCA+SVM (Strategy D) at 82.61% (size 5). Across most settings, accuracies remain stable with differences rarely exceeding 2 pp. Only extreme compression (1–2 components) leads to noticeable degradation. SVM consistently performs slightly better than kNN, particularly with NCA, suggesting superior generalization from the transformed feature space.

2-class setting. The best accuracy is 90.63% with NCA+kNN (Strategy C, 640 components). NCA+SVM (Strategy A) is a close second at 90.62% with only 5 components. Multiple configurations exceed 90.4–90.5% across wide compression ranges. At extreme compression (1–2 components), performance generally remains close to mid-range results, underscoring robustness. The main exception is NCA+SVM (Strategy A) at size 1 (72.34%).

Comparison to baselines. Table 3 summarizes the best configurations for both settings alongside historical baselines. The top 4-class result (82.70%) exceeds the strongest reported baseline (VGG19 [1], 81.62%) by more than one percentage point. Other top configurations include NCA+kNN (82.63%) and PCA+SVM (82.61%), demonstrating consistent gains across compression ratios from 4× to over 60×. In the binary setting, nearly all configurations outperform historical baselines, with compressed feature spaces of 5–50 components retaining or surpassing full-dimensional performance.

Cross-dataset validation on BreakHis. Table 4 reports, for each strategy and magnification, the best accuracy over all compression sizes (main value) and the result with only two components (parenthesized); per-column maxima are in **bold** (best overall) and underlined (best at 2 components).

BreakHis accuracies are tightly clustered at 93–96%. Among PCA-based pipelines, PCA+SVM (Strategy A) attains 95.94% at 40×, while PCA+kNN (Strategy C) peaks at 94.86% for 100×. NCA+kNN (Strategy A) leads at 400× with 94.78%.

The parenthesized 2-component accuracies serve as a compression stress test. NCA-based pipelines retain ∼88–91% across magnifications (e.g., 91.48% for NCA+kNN Base at 40×; 90.68% at 100×), substantially outperforming PCA-based pipelines (66–81%). Among strategies, Base and A frequently yield the highest compressed-feature values, while C and D show more variability.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_31

Choice of refinement strategy. Strategy A emerges as the top performer in three of four DiagSet pipelines (Table 3). From a modeling perspective, it concatenates head with Stage 4 features – representations closest to the decision layer – capturing high-level semantics while avoiding noise from earlier stages. We therefore adopt Strategy A for all refinement ablations, applying it post hoc on identical predictions (no retraining).

3.2 Spatial refinement results

Using re-extracted patches with spatial coordinates, we compare the baseline to post-hoc refinement on the exact same model outputs (identical weights, inputs, inference). Two consistent patterns emerge from Table 5.

First, refinement improves both accuracy and F1 in every pipeline at both tested dimensionalities (640 and 5 components). PCA+kNN improves from 90.17% to 92.16% (F1 score from 0.895 to 0.904) at 640 components, and from 90.23% to 91.94% (F1 score from 0.905 to 0.909) at 5 components. NCA+kNN rises from 90.48% to 92.27% (F1 score from 0.899 to 0.911) at 5 components. SVM-based pipelines show similar but slightly smaller gains (e.g., NCA+SVM: 90.00% to 91.45% at 5 components).

Second, gains are observed at both high and low dimensionalities, suggesting the refinement effect is largely independent of compression level: kNN benefits by ≈1.6–2.0 pp and SVM by ≈0.8–1.7 pp. Accuracy and F1 move in tandem, with the largest improvements for kNN (e.g., PCA+kNN at 640 components: +1.99 pp accuracy, +0.009 F1).

Table 1: Classification accuracy (%) on DiagSet, 4-class setting across pipelines. Best per pipeline in **bold**; best overall underlined.

	kNN				SVM			
# comp	PCA Base	NCA Base	PCA Str. A	NCA Str. A	PCA Base	NCA Base	PCA Str. A	NCA Str. A
960	–	–	82.54	82.64	–	–	82.09	82.18
640	–	–	82.54	82.43	–	–	82.13	82.12
320	82.23	82.39	82.54	82.53	82.15	82.08	81.98	82.04
160	82.35	82.37	82.55	**82.65**	81.69	81.69	81.64	81.70
80	82.40	**82.63**	82.37	82.47	81.36	81.38	81.80	82.08
40	82.39	82.54	82.41	82.32	82.11	82.07	82.34	82.22
20	82.42	82.35	82.54	82.59	82.20	82.26	**82.42**	**82.34**
10	**82.46**	82.17	82.53	82.15	**82.41**	81.93	82.28	81.85
5	82.35	82.01	82.33	82.07	82.36	**82.53**	82.27	82.12
2	82.31	81.90	82.10	81.67	82.11	82.05	81.99	81.82
1	81.79	81.06	80.82	81.25	81.73	81.79	80.87	81.97

Only Base and Strategy A shown for clarity. Strategies B–D follow similar trends; best results: Str. B 82.48 (NCA-kNN), Str. C 82.32 (PCA-SVM), Str. D 82.70 (NCA-SVM). Full per-strategy tables available in supplementary material in repository.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_31

Table 2: Classification accuracy (%) on DiagSet, 2-class setting. Best per pipeline in **bold**; best overall underlined.

	kNN				SVM			
# comp	PCA Base	NCA Base	PCA Str. A	NCA Str. A	PCA Base	NCA Base	PCA Str. A	NCA Str. A
960	–	–	**90.49**	90.33	–	–	89.78	89.76
640	–	–	**90.49**	90.16	–	–	89.58	89.44
320	90.32	90.33	90.43	90.26	90.01	90.04	89.98	90.01
160	90.31	**90.45**	90.38	90.31	90.20	90.18	90.12	90.08
80	**90.44**	90.43	90.34	90.35	90.41	90.37	90.32	90.34
50	90.35	90.43	90.43	90.37	**90.47**	**90.54**	90.45	90.47
30	90.38	90.10	90.34	90.34	90.41	90.43	90.46	90.46
20	90.41	90.28	90.38	**90.38**	90.41	90.43	90.42	90.33
5	90.36	90.38	90.19	90.12	90.43	90.30	**90.49**	**90.62**
2	90.26	89.64	90.28	90.04	90.26	90.15	90.40	90.42
1	90.16	90.09	90.45	89.84	90.32	90.12	90.38	72.34

Only Base and Strategy A shown. Best Str. B–D: B 90.45 (PCA-kNN), C 90.63 (NCA-kNN), D 90.45 (PCA-SVM).

Table 3: Best results per pipeline on DiagSet (4-class and 2-class) compared with prior work.

	4-class			2-class		
Pipeline	Str.	#	Acc.	Str.	#	Acc.
PCA-kNN	Base	10	82.46	Base	80	90.44
PCA-kNN	Str. A	270	82.56	Str. A	640	90.49
PCA-SVM	Base	10	82.41	Base	50	90.47
PCA-SVM	Str. D	5	82.61	Str. A	5	90.49
NCA-kNN	Base	80	82.63	Base	160	90.45
NCA-kNN	Str. A	160	82.65	Str. C	640	**90.63**
NCA-SVM	Base	5	82.53	Base	50	90.54
NCA-SVM	Str. D	30	**82.70**	Str. A	5	90.62
EffNet-B4 [10]	–	–	81.35	–	–	90.37
VGG19 [1]	–	–	81.62	–	–	90.24

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_31

Table 4: Best accuracy (%) on BreakHis per strategy and magnification. Parenthesized: accuracy with 2 components only. **Bold**: best overall; underlined: best at 2 comp.

Setting	Str.	40×	100×	200×	400×
PCA-kNN	Base	95.89 (76.94)	94.52 (80.68)	**94.59** (77.30)	94.18 (76.15)
	A	95.79 (71.88)	94.38 (76.93)	94.54 (74.62)	94.12 (73.79)
	C	95.64 (69.12)	**94.86** (71.26)	94.34 (68.60)	94.18 (66.70)
PCA-SVM	A	**95.94** (72.68)	94.52 (77.61)	94.24 (75.91)	93.13 (72.20)
NCA-kNN	Base	95.79 (91.48)	94.57 (90.68)	94.39 (89.07)	94.12 (89.18)
	A	95.74 (89.57)	94.52 (89.33)	94.24 (89.27)	**94.78** (88.02)
NCA-SVM	Base	95.74 (91.48)	94.47 (89.67)	94.19 (89.32)	93.41 (88.96)
	A	95.84 (89.67)	94.43 (89.57)	94.29 (89.57)	93.30 (87.25)

Strategies B, C (PCA-SVM), D omitted; trends similar with slightly lower peaks.

Table 5: Ablation of post-hoc refinement on DiagSet (Strategy A): accuracy (%) and F1 score at two different embedding sizes.

		Original		Refined	
Setting	# comp	Acc.	F1	Acc.	F1
PCA-kNN	640	90.17	0.895	92.16	0.904
	5	90.23	0.905	91.94	0.909
PCA-SVM	640	88.59	0.841	89.42	0.846
	5	90.24	0.902	91.92	0.911
NCA-kNN	640	90.20	0.904	91.96	0.906
	5	90.48	0.899	92.27	0.911
NCA-SVM	640	88.37	0.839	89.24	0.852
	5	90.00	0.888	91.45	0.892

4 Discussion

4.1 Base pipeline

Across DiagSet-A.1, compressed TinyViT features achieve strong accuracy in both settings. The best 4-class configuration (NCA+SVM, Strategy D) reaches 82.70%, while multiple pipelines exceed 90.4–90.6% in the binary task. Comparing Base and concatenation strategies indicates that late-stage representations are particularly useful. Strategy A (head + Stage 4) performs as a non-inferior, stable choice, often matching per-block maxima while achieving top performance at very low dimensionalities (5–20 components).

Replicating on BreakHis, no single strategy dominates across all blocks; however, when comparing projection methods under comparable compression (2

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_31

components), NCA consistently outperforms PCA for both kNN and SVM. The primary cross-dataset signal is thus a projection effect (NCA>PCA) rather than a decisive ranking among strategies. We observe complementary benefits: PCA's unsupervised projection preserves global variance and supports aggressive compression with SVM, while NCA's supervised transform better preserves class neighborhoods, explaining its strong 2-component performance.

4.2 Refinement ablation

The post-hoc ablation (identical weights, no retraining) with threshold τ=0.99 consistently improves both accuracy and F1 across all pipelines and embedding sizes. Gains at both 640 and 5 dimensions suggest the effect is largely independent of compression level. This indicates that a lightweight post-hoc adjustment can deliver reliable benefits without modifying the backbone or training recipe. Beyond accuracy, the resulting confidence and uncertainty maps offer structured spatial representations suitable for downstream integration with vision–language models.

4.3 Future work

We plan to (i) extend refinement to multi-class settings with per-class PR/ROC analysis; (ii) evaluate patient-level aggregation and cost-aware metrics (throughput, latency, memory); (iii) leverage the structured uncertainty maps as spatial priors for large vision–language models that integrate image evidence with textual diagnostic reasoning.

5 Conclusions

We investigated compact representations from TinyViT-5M for histopathology, combining five feature strategies with PCA/NCA projection and kNN/SVM classifiers. On DiagSet-A.1, the best 4-class accuracy reaches 82.70% (NCA+SVM, Strategy D) and several binary configurations exceed 90.5%. Cross-dataset replication on BreakHis yields tightly clustered accuracies (93–96%), with NCA consistently outperforming PCA under aggressive compression (2-component stress test: ∼88–91% vs. 66–81%).

A controlled post-hoc refinement ablation shows consistent gains in both accuracy (+0.8–2.0 pp) and F1 across projection methods, classifiers, and embedding sizes (640 and 5), indicating independence from compression level. Taken together, these results demonstrate that (i) TinyViT features maintain high accuracy across compression levels, (ii) supervised dimensionality reduction (NCA) is preferable under tight compression, and (iii) post-hoc spatial refinement provides consistent, low-cost improvements on fixed outputs. Future work will consider slide/patient-level endpoints and exploit structured uncertainty maps for vision–language models.

The implementation is available at: https://github.com/niedziel96/TinyViT-Refine

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_31

Acknowledgments. This research was funded in whole by the National Science Centre, Poland, grant number: UMO-2024/55/B/ST6/01681.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Koziarski, M., Cyganek, B., Niedziela, P., Olborski, B., Antosz, Z., Żydak, M., et al.: DiagSet: A dataset for prostate cancer histopathological image classification. Sci. Rep. 14(1), 6780 (2024)
2. Wu, K., Zhang, J., Peng, H., Liu, M., Xiao, B., Fu, J., Yuan, L.: TinyViT: Fast pretraining distillation for small vision transformers. In: ECCV, pp. 68–85. Springer (2022)
3. Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv:2010.11929 (2020)
4. DeepSeek-AI: DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning (2025)
5. Vorontsov, E., et al.: Virchow: A million-slide digital pathology foundation model. arXiv:2309.07778 (2024)
6. Zimmermann, E., et al.: Virchow2: Scaling self-supervised mixed magnification models in pathology (2024)
7. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: ICCV, pp. 9650–9660 (2021)
8. Oquab, M., et al.: DINOv2: Learning robust visual features without supervision. arXiv:2304.07193 (2023)
9. Amir, S., Gandelsman, Y., Bagon, S., Dekel, T.: Deep ViT features as dense visual descriptors. arXiv:2112.05814 (2021)
10. Alici-Karaca, D., Akay, B.: An efficient deep learning model for prostate cancer diagnosis. IEEE Access (2024)
11. Cui, B., Islam, M., Bai, L., Ren, H.: Surgical-DINO: Adapter learning of foundation models for depth estimation in endoscopic surgery. IJCARS, pp. 1–8 (2024)
12. Müller-Franzes, G., et al.: Medical Slice Transformer: Improved diagnosis and explainability on 3D medical images with DINOv2. arXiv:2411.15802 (2024)
13. Goldberger, J., Hinton, G., Roweis, S., Salakhutdinov, R.: Neighbourhood components analysis. In: NeurIPS, vol. 17, pp. 513–520 (2005)
14. Spanhol, F., Oliveira, L.S., Petitjean, C., Heutte, L.: A dataset for breast cancer histopathological image classification. IEEE TBME 63(7), 1455–1462 (2016)
15. Crow, F.C.: Summed-area tables for texture mapping. In: SIGGRAPH, pp. 207–212. ACM (1984)
16. Valanarasu, J.M.J., et al.: Medical Transformer: Gated axial-attention for medical image segmentation. arXiv:2102.10662 (2021)
17. Chen, J., et al.: TransUNet: Transformers make strong encoders for medical image segmentation. arXiv:2102.04306 (2021)
18. Łażewski, S., Cyganek, B.: Highly compressed image representation for classification and content retrieval. Integr. Comput.-Aided Eng. 31(3), 267–284 (2023)
19. Knapik, M., Cyganek, B.: Fast eyes detection in thermal images. Multimed. Tools Appl. 80, 3601–3621 (2021)

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_31