# Quo vadis pathology? Advancing glomerular lesion classification with foundation models

David Lima*§, Grinaldo Oliveira*§, Ângelo Duarte†, Washington Santos‡ and Luciano Oliveira*‡
*Intelligent Vision Research Lab, Institute of Computing, Federal University of Bahia
Email: davidlima,grinaldooliveira,lrebouca{@ufba.br}
†High-Performance Computing Lab, Department of Technology, State University of Feira de Santana
Email: angeloduarte@uefs.br
‡Structural and Molecular Pathology Lab, Gonçalo Moniz Institute, Fundação Oswaldo Cruz
Email: wluis@bahia.fiocruz.br

2025 38th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI) | 979-8-3315-8951-6/25/$31.00 ©2025 IEEE | DOI: 10.1109/SIBGRAPI67909.2025.11223393

***Abstract*—Computational pathology is undergoing a significant transformation with the emergence of foundation models (FMs). These models leverage self-supervised learning on extensive histopathological datasets with the aim of extracting robust feature representations. FMs hold potential to automate advanced diagnostic pipelines, encompassing segmentation, classification, and biomarker discovery. This study evaluates the effectiveness of embeddings from four SOTA FMs (UNI, UNI2, Phikon, and Phikon2) for one-versus-all glomerular lesion classification. We propose here a comparative framework in which a multilayer perceptron (MLP) and a support vector machine (SVM) – each trained exclusively on FM-derived embeddings – are benchmarked against EfficientNet, a fully supervised end-to-end image classifier. By varying the number of cross-validation folds (from k=2, representing minimal training data, to k=5, representing maximal training data), on a proprietary histopathology dataset, we assess classifier robustness under differing data regimes. Our results demonstrate that, even without any FM fine-tuning, the UNI/SVM pipeline outperforms the EfficientNet by 3.4 percentage points in average F1-score, considering all values of k.**

## I. Introduction

The glomerulus is a renal structure capable of filtering blood [1]. Due to its primary filtering nature in the kidney, pathologists always check its integrity during the diagnosis process to identify kidney diseases and systemic conditions with the potential to affect the kidney [2], making it a particularly reliable diagnostic indicator.

Diagnosing glomerular lesions is often a laborious and specialized task that relies on a histological examination performed by experienced professionals, who do not always reach a consensus [3]. Thus, researchers started to develop intelligent systems that provide support in the acquisition, management, and interpretation of pathology information, which culminated in the storage of a large amount of pathological data digitally [3]. This led to an increase in the use of image feature extraction and machine learning classification approaches, such as KNN techniques [4] and CNN-based architectures: InceptionV3 [5], Xception [6], VGG-19 [2], ResNet50 [7] and other proposed architectures [2], [8], [9]. Indeed, most existing approaches rely on supervised training, which requires a specialized dataset and a large amount of labeled data.

Recently, foundation models, large models pre-trained on various datasets without task-specific labels, offer a promising alternative by learning general visual representations that can be transferred to downstream applications with minimal to no fine-tuning [10]. Also, researchers have already developed FMs specialized in histopathological data [11]–[13]. Although some combine convolutional and transformer architectures to capture hierarchical tissue patterns [14], most are based on a Vision Transformer (ViT) network [15]. Overall, these models have shown promising results in slide embedding and tumor classification [16], but their utility for fine-grained glomerulus lesion classification remains unexplored.

In this study, we investigated whether foundation models can achieve competitive performance without fine-tuning. The methodology involved using embeddings from four state-of-the-art FMs (UNI, UNI2, Phikon, and PhikonV2) within a glomerular lesion classification pipeline. These embeddings were used to train MLP and SVM classifiers on a proprietary dataset containing 7,566 glomerulus images across six lesion classes. Compared to an EfficientNet-B0 baseline model, SVM, using embeddings from UNI and UNI2, demonstrated superior performance, achieving an average improvement of 3.4 F1-score percentage points over the baseline and greater consistency in results (indicated by generally smaller standard deviations). This finding validates the central hypothesis that the rich, generalizable features learned by large-scale FMs are highly discriminative and relevant for glomerular lesion classification, even without extensive fine-tuning.

## II. Related work

Glomerular lesions impair kidney filtration and are critical indicators of renal disease. The variability in morphology and staining remains a major challenge for automated classification [8].

Early approaches relied on handcrafted features and classical machine learning. Barros et al. [4], for example, combined image pre-processing with a k-nearest neighbor classifier to identify hypercellularity lesions, reaching 88.3% precision but

§Equal contribution.

979-8-3315-8951-6/25/$31.00 ©2025 IEEE

TABLE I
SUMMARY OF THE FOUNDATION MODELS USED IN THIS STUDY, INCLUDING ARCHITECTURE, PRE-TRAINING DATASET SIZE, ACCOMPLISHED TASKS, ROBUSTNESS INDEX, AND AVAILABILITY OF MODEL WEIGHTS.

| **Model** | **PT dataset size** | **Backbone** | **Primary tasks** | **Robustness index** [17] | **Weight availability** |
|---|---|---|---|---|---|
| **UNI** [11] | >100M patches, 100K WSIs | ViT-L/16 | General pathological image interpretation, disease detection, diagnosis, transplant assessment | 0.88 | Restricted access |
| **UNI2** [11] | >200M tiles, 350K H&E/IHC slides | ViT-H/14 | ROI/slide classification, segmentation, feature extraction | 0.93 | Restricted access |
| **Phikon** [12] | ∼460M tiles (public) | ViT-B/16 | Feature extractor for biomarkers, histological analysis | 0.84 | Public |
| **PhikonV2** [13] | 450M images (public) | ViT-L/16 | Feature extractor for biomarkers, ROI/slide classification, segmentation | 0.74 | Public |

remaining limited by feature design. Deep learning methods soon advanced the field: Chagas et al. [8] proposed a hybrid CNN-SVM framework that achieved near-perfect binary classification (F1 = 99.6%) and competitive multi-class performance (F1 = 82%), demonstrating the linear separability of CNN features and surpassing both traditional pipelines and state-of-the-art CNNs (Xception, ResNet50, InceptionV3).

Further, Barros et al. [2] introduced PodNet for podocytopathy detection, leveraging VGG19 feature extraction across multiple color spaces (RGB, HED, HDX). The approach achieved an F1-score of 90.9%, highlighting the benefit of complementary representations for rare glomerular diseases.

Beyond nephropathology, foundation models (FMs) have shown promise in medical image analysis. Li et al. [18] benchmarked embeddings from MedImageInsight for tube placement in radiographs, where SVMs trained on FM features reached an mAUC of 93.8%. Similarly, Enda et al. [19] demonstrated that linear probing with the UNI model — specialized for pathology images — outperformed ImageNet-pretrained ViTs, CTransPath [14], and full fine-tuning for brain tumor classification, underscoring the adaptability of pathology-focused FMs.

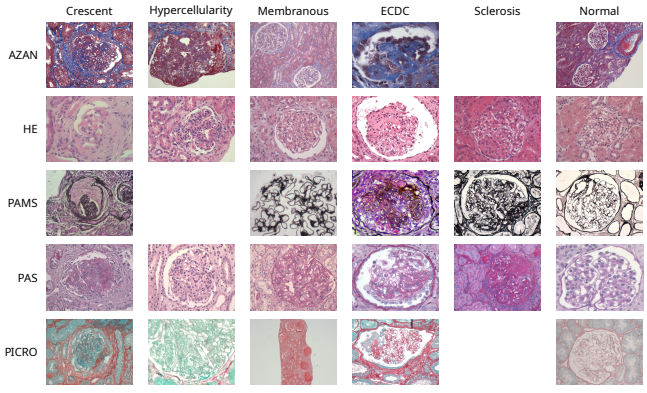



Fig. 1. Sample images from our dataset. Blank cells indicate class–staining combinations not available in the collection.

## III. FMs FOR HISTOPATHOLOGY

The existence of a plethora of FMs within computational pathology is primarily attributed to the inherent complexity of digital pathological data and the broad spectrum of clinical and research applications sought within this field [20]. Histopathological images, particularly whole slide images (WSIs), are inherently complex due to their gigapixel resolution, variability in staining and tissue preparation, analysis across multiple magnifications, and contextual fragmentation introduced by tiling. Addressing these complexities demands models with architectural flexibility and processing strategies capable of handling heterogeneous data, effectively [2], [9], [16], [21].

Computer vision FMs incorporate a wide range of neural architectures, including ViTs [15], CNNs [22], Swin Transformers [23], and others [14]. The primary strategy adopted to train these models is supervised learning frameworks, such as DINO [24], DINOv2 [25], iBot [26], SimCLR [27], MoCo [28], CLIP [29], and CoCa [30]. These frameworks emphasize different aspects of representation learning, aiming to improve model robustness, scalability, and generalization.

FMs also vary substantially in scale, both in terms of model capacity (number of parameters) and the diversity and size of their pretraining datasets, where in the context of histopathology, ranges from thousands to millions of WSIs or tiles [12], [16], [20]. Computational pathology FMs are developed with adaptability in mind, targeting a broad spectrum of downstream tasks. These include patch-level and slide-level classification (*e.g.*, pan-cancer detection and subtyping), segmentation and detection of cells or lesions, content-based image retrieval, report generation, visual question answering, and predictive tasks such as biomarker status estimation or prognostic modeling [11], [31].

Particularly in our work, the FMs selected – UNI [11], UNI2 [11], Phikon [12], and PhikonV2 [13] – represent cutting-edge efforts in developing large-scale and generalizable AI for histopathology, and based primarily on ViTs. Table I summarizes the main characteristics of these FMs.

### A. UNI and UNI2

UNI was developed by researchers at Harvard Medical School, constituting a pioneering FM designed specifically to

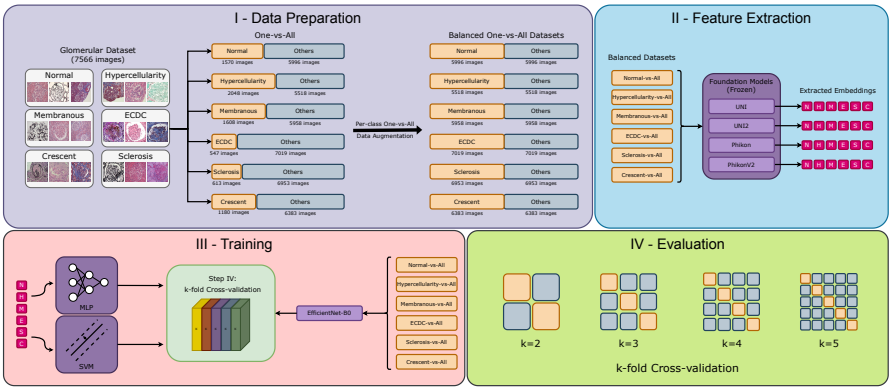



Fig. 2. Experimental protocol. (I) Each lesion class was separated one-vs-all and balanced via augmentation, producing six class-specific datasets. (II) Four foundation models (UNI, UNI2, Phikon, PhikonV2) extracted latent representations, yielding 24 image sets. (III) Consistency was ensured through cross-validation, contrasting raw histopathology (EfficientNet-B0) with transfer learning (FM+MLP/SVM). (IV) Classifiers were evaluated per lesion class (e.g., crescent vs. all) using k-fold cross-validation ($k \in [2, 5]$) to test robustness and generalizability.

interpret pathological images [11]. It is capable of recognizing disease in both specific regions of interest and gigapixel whole-slide images. UNI was trained on a dataset, comprising over 100 million tissue patches and more than 100,000 whole-slide images. Through extensive evaluation, UNI demonstrated strong performance across 34 diverse diagnostic tasks, including various cancer classifications and organ transplant assessment. UNI2 is a successor to the UNI model, pretrained on over 200 million image patches derived from more than 350,000 H&E and IHC slides sourced from Mass General Brigham. It provides precomputed embeddings for widely used datasets such as TCGA, CPTAC, and PANDA. It supports region-of-interest (ROI) classification via linear probing and slide-level classification through multiple instance learning. A key improvement over its predecessor is a reported robustness index of 0.93 compared to 0.88 for UNI [17].

### B. Phikon and PhikonV2

Phikon is a ViT-based model developed for histological image analysis, demonstrating strong robustness to confounding factors such as medical center of origin, with a reported robustness index of 0.84 [17]. PhikonV2 builds on Phikon with substantial improvements in scale and performance. It is a ViT-Large architecture with approximately 0.3 billion parameters, pretrained using self-supervised learning with masked image modeling and multi-crop loss via the DINOv2 framework. The model was trained on 450 million histological image patches from over 100 public cohorts spanning more than 30 cancer types. PhikonV2 supports various downstream tasks, including ROI classification, slide-level classification, and segmentation. It presents lower robustness compared to Phikon (0.74 vs. 0.84).

## IV. Experimental Analysis

### A. Dataset

As illustrated in Fig. 1, our dataset comprises six glomerular lesion classes (normal, hypercellularity, membranous, epithelial cell degenerative changes (ECDC), crescentic, and sclerotic) captured under multiple staining protocols. To assess clinical applicability and model generalization, we curated a proprietary set of 7,566 glomerulus images, each independently annotated by four pathologists from two different countries. The samples were stained using Hematoxylin and Eosin (HE), Periodic Acid–Schiff (PAS), Picro–Sirius Red (PICRO), Azocarmine and Aniline Blue (AZAN), and Periodic Acid Methenamine Silver (PAMS). Certain lesions were not stained with every technique, resulting in missing class–staining combinations.

Employing a non-public image source is a deliberate methodological choice, as foundation models are typically pretrained on large-scale, publicly aggregated, or consortium-specific collections.

### B. Evaluation protocol

The central objective of this study is to empirically determine the effectiveness of using embeddings derived from SOTA FMs in computational pathology. The hypothesis is that the rich, generalizable features learned and encapsulated in the pre-trained embeddings of these large-scale FMs can be effectively leveraged by computationally simpler downstream classifiers, offering an accurate alternative to training complex end-to-end models from scratch on limited and task-specific annotated data.

As illustrated in Fig. 2, we used a multi-step pipeline designed to compare FMs embeddings' separability. **First (Fig. 2-I)**, we rearranged the main dataset in a one-vs-all configuration, transforming the multiclass task into multiple binary classification problems. Due to their one-vs-all nature, these datasets were significantly unbalanced, presenting the need to balance their classes. After applying a variety of image data augmentation techniques (random vertical/horizontal flips, affine transforms, gaussian noise/blur, color jitter, coarse dropout, grid distortion, piece-wise affine transform and optical distortion) to each dataset, we ended up with six different one-vs-all datasets. **In the second step (Fig. 2-II)**, we introduced our FMs to the pipeline by using each to

TABLE II

PER-$k$ AVERAGE F1-SCORE ACHIEVED BY EACH MODEL THROUGH CROSS-VALIDATION. HIGHLIGHTED VALUES WERE CHOSEN BASED ON THE HIGHEST AVERAGE, THE LOWEST STANDARD DEVIATION ($\sigma$), AND – FOR CLOSELY COMPETING VALUES – THE LOWEST DISPERSION COEFFICIENT.

| | | | SVM | | | | MLP | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **#Folds ($k$)** | **Class** | **EfficientNet-B0** | **UNI** | **UNI2** | **Phikon** | **PhikonV2** | **UNI** | **UNI2** | **Phikon** | **PhikonV2** |
| 2 | Hypercellularity | 86.1±0.0% | **92.2±0.4%** | 91.2±0.2% | 24.9±18.2% | 23.4±19.6% | 89.4±0.9% | 88.3±0.8% | 42.4±0.4% | 42.7±0.3% |
| | Normal | 90.9±1.2% | **93.8±0.3%** | 93.3±0.7% | 20.4±14.6% | 19.9±14.7% | 91.2±0.4% | 92.1±1.0% | 23.0±11.9% | 20.7±13.7% |
| | Membranous | 87.6±0.9% | **92.9±0.2%** | 91.8±0.2% | 36.4±0.7% | 36.3±0.7% | 90.0±0.3% | 88.7±0.0% | 36.8±0.7% | 36.4±0.7% |
| | Sclerosis | 84.7±1.7% | **86.9±0.0%** | 86.1±0.3% | 19.6±0.8% | 13.4±3.4% | 82.2±1.4% | 80.6±1.5% | 17.1±1.0% | 15.9±1.2% |
| | Crescent | 88.5±0.8% | **91.4±0.3%** | 91.1±1.5% | 27.6±0.1% | 27.4±0.1% | 88.7±1.1% | 88.8±0.6% | 27.8±0.0% | 27.5±0.0% |
| | ECDC | 90.9±0.5% | **93.3±0.4%** | 93.3±0.5% | 16.1±2.7% | 14.1±4.3% | 89.6±0.6% | 88.8±0.3% | 20.9±0.0% | 15.4±2.5% |
| 3 | Hypercellularity | 86.4±2.2% | **93.4±0.4%** | 92.9±0.3% | 31.1±16.7% | 30.2±17.6% | 91.0±1.1% | 90.6±0.2% | 30.8±17.2% | 42.4±0.2% |
| | Normal | 92.8±1.4% | **95.1±0.5%** | 94.3±0.4% | 35.2±0.2% | 35.0±0.4% | 92.6±0.3% | 92.2±0.8% | 26.3±12.0% | 35.1±0.4% |
| | Membranous | 88.5±1.1% | **94.1±0.9%** | 93.8±0.3% | 36.4±1.8% | 36.3±1.8% | 91.6±1.3% | 92.2±0.9% | 36.5±1.7% | 36.6±2.0% |
| | Sclerosis | 87.5±2.5% | 89.6±1.9% | **89.8±0.6%** | 20.0±3.3% | 12.4±3.0% | 85.2±2.2% | 83.6±1.6% | 19.4±3.6% | 18.3±3.9% |
| | Crescent | 88.7±1.9% | **92.6±1.1%** | 92.3±1.3% | 27.7±0.6% | 27.4±0.7% | 90.3±0.5% | 89.7±1.9% | 27.8±0.6% | 27.6±0.7% |
| | ECDC | 93.7±1.4% | 94.5±0.9% | **94.5±0.8%** | 21.0±3.9% | 12.2±3.8% | 90.0±1.2% | 91.1±0.5% | 23.5±2.9% | 18.3±3.4% |
| 4 | Hypercellularity | 88.7±1.4% | **94.1±1.0%** | 93.3±0.6% | 34.4±14.9% | 33.6±15.9% | 91.8±1.5% | 91.1±1.3% | 42.4±1.0% | 33.7±14.6% |
| | Normal | 93.3±1.3% | **95.5±0.7%** | 94.7±0.4% | 35.2±0.7% | 35.1±0.8% | 93.1±0.4% | 93.0±0.7% | 35.2±0.7% | 27.8±12.7% |
| | Membranous | 88.5±1.8% | **94.8±0.9%** | 94.3±0.3% | 36.4±0.9% | 36.3±0.9% | 92.3±1.5% | 92.1±0.2% | 36.8±1.0% | 36.6±0.9% |
| | Sclerosis | 87.9±4.8% | 90.0±1.5% | **90.3±0.9%** | 19.7±2.3% | 12.3±3.8% | 86.1±0.8% | 85.4±0.9% | 17.4±3.6% | 18.1±2.9% |
| | Crescent | 89.2±1.7% | **93.1±0.6%** | 93.1±0.8% | 27.8±0.8% | 27.4±0.9% | 90.5±0.8% | 90.7±1.0% | 27.9±0.8% | 27.4±1.1% |
| | ECDC | 91.7±3.0% | **94.6±1.1%** | 94.5±1.1% | 20.8±4.3% | 10.3±4.4% | 90.4±1.6% | 90.5±1.9% | 22.3±4.3% | 19.5±2.7% |
| 5 | Hypercellularity | 89.9±0.9% | **94.1±1.2%** | 93.6±0.8% | 43.0±0.9% | 42.7±0.7% | 91.0±0.7% | 91.3±1.0% | 29.3±16.1% | 36.0±13.0% |
| | Normal | 94.0±0.9% | **95.7±1.1%** | 95.3±0.5% | 35.2±1.5% | 35.0±1.7% | 93.6±1.2% | 93.5±1.0% | 35.2±1.6% | 35.0±1.6% |
| | Membranous | 91.1±2.2% | **94.9±1.1%** | 94.5±0.9% | 36.4±1.5% | 36.3±1.7% | 92.8±0.6% | 92.5±0.4% | 36.7±1.4% | 36.5±1.5% |
| | Sclerosis | 89.2±2.8% | 90.8±1.3% | **91.1±1.3%** | 19.9±5.1% | 12.5±3.4% | 87.3±0.8% | 86.1±1.7% | 19.0±4.7% | 17.3±3.6% |
| | Crescent | 91.0±1.1% | **93.6±0.9%** | 93.3±1.0% | 27.8±1.0% | 27.4±1.0% | 91.0±1.0% | 90.7±1.1% | 27.8±1.0% | 27.6±1.0% |
| | ECDC | 93.4±1.1% | 94.9±0.8% | **95.0±0.8%** | 20.7±5.1% | 10.7±3.2% | 90.7±1.4% | 90.8±1.3% | 23.3±3.4% | 17.4±4.7% |
| **Avg difference to baseline ($\Delta$)** | | +0.0pp | **+3.4pp** | +3.04pp | -61.69pp | -64.44pp | 0.34pp | 0.01pp | -61.19pp | -61.85pp |

TABLE III

SUMMARY OF COMPUTATIONAL COST AND RUNTIME FOR EACH MODEL.

| **Model** | **Avg GPU memory (MB)** | **Prediction time (ms)** |
|---|---|---|
| EfficientNet-B0 | 355.9 | 18.58 ± 3.16 |
| UNI | 1,491.0 | 12.37 ± 1.95 |
| UNI2 | 3,392.0 | 18.60 ± 1.04 |
| Phikon | 606.6 | 12.05 ± 1.23 |
| PhikonV2 | 1,538.7 | 14.56 ± 1.99 |

generate embeddings of every dataset produced in the previous step. Repeating this procedure for every dataset and FM, we concluded this step with 24 sets of embeddings, six produced by each FM. **Finally (Figs. 2-III and -IV)**, we evaluated the representativeness of the embeddings produced by each FM by training two classifiers (MLP and SVM) over them in a varying $k$-fold cross-validation fashion, to check the robustness of the obtained results across an increasing number of training images ($k \in [2, 5]$). The decision to repeat the $k$-fold cross-validation step with different values of $k$ derives from the need to evaluate the robustness of the models as the number of training images ($\frac{k-1}{k}$) increases and test images ($\frac{1}{k}$) decreases. Repeating this procedure with increasing values of $k$ gives us important insight in how the models behave when different amounts of training data are available for the top-layer training. In our experiments, we varied $k$ for values in the interval $[2, 5]$, going from $50\%$ to $80\%$ of the dataset dedicated to training.

For each FM classifier, hyperparameter optimization was performed using a grid search strategy aimed at maximizing predictive performance. For the MLP, we employ a compact architecture with a single hidden layer of 100 ReLU units. To curb overfitting, we use $\alpha = 0.01$ weight decay, early stopping with a 10% validation split, and patience of 10 epochs. Its training is performed with the Adam optimizer (constant learning rate, $\text{lr}_{\text{init}} = 10^{-3}$) over a maximum of 100 iterations, and a convergence tolerance of $10^{-3}$. For the SVM model, a radial basis function (RBF) kernel was used with a moderately high regularization parameter (*C=10.0*) in order to balance model complexity. To improve generalization, the RBF-kernel bandwidth parameter $\gamma$ was scaled adaptively to the variance of the input embeddings, given by

$$\gamma = \frac{1}{N \cdot \text{Var}(\mathbf{X})}, \tag{1}$$

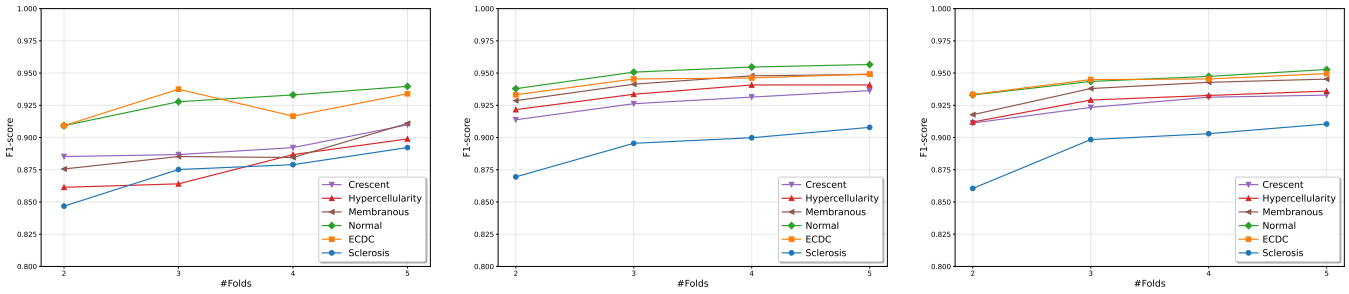
where $N$ is the number of features and Var(X) is the variance of the set containing all training samples.

For a baseline, we used the same training configuration to finetune an EfficientNet-B0 model directly over the images produced in the first step. It is noteworthy that every model shares the same folds – each fold with the same images – during the varying $k$-fold cross-validation procedure. During validation, samples derived from augmentation were not included in the validation fold; instead, they were confined to the training folds. This was necessary to assess the classifier's performance using original samples only.

### C. Result Analysis

As shown in Table II, the SVM classifiers trained on UNI and UNI2 embeddings significantly outperformed the EfficientNet-B0 baseline, achieving average F1-score gains ($\Delta$) of +3.40 percentage points (UNI+SVM) and +3.05 percentage points (UNI2+SVM). The value of $\Delta$ is found as

$$\Delta(X(m)) = \frac{\sum_{k\in[2,5], c\in C} (x_k^c - b_k^c)}{|X(m)|}, \tag{2}$$




(a) EfficientNet-B0 (b) UNI + SVM (c) UNI2 + SVM

Fig. 3. Average F1-scores per $k$ from $k$-fold cross-validation on each balanced one-vs-all dataset: (a) fine-tuned EfficientNet-B0, (b) UNI + SVM (best-performing FM), and (c) UNI2 + SVM (alternative FM).

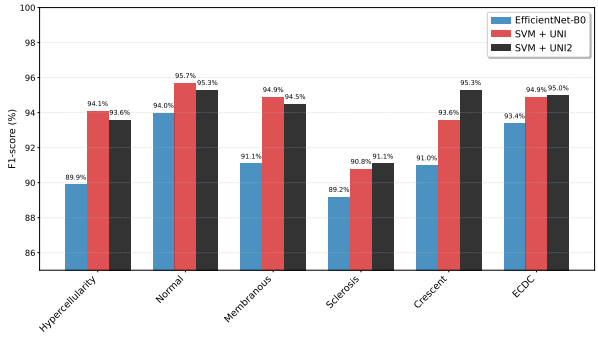



Fig. 4. Evaluation of top-performing classifiers using 5-fold cross-validation.

where $X(m)$ is the set of all results achieved by a particular model $m$, $C$ is the set containing all classes $c$ evaluated, $B$ is the results achieved by the baseline, $x_k^c \in X(m_1)$ and $b_k^c \in X(m_2)$ are the average F1-scores achieved by a model in $k$-fold cross-validation in a $c$-vs-all setting, for $k \in [2, 5]$.

The smaller $\sigma$ values observed for SVM-UNI and SVM-UNI2 (average of 0.81 and 0.69, respectively) indicate greater performance consistency and stability across folds, underscoring the robustness of their embeddings for glomerular lesion classification. UNI and UNI2 were explicitly trained on pathology-oriented datasets, which likely enhances their ability to capture morphological and textural patterns relevant to glomeruli. Their stable behavior across classes further suggests that such embeddings generalize well even under class imbalance and staining variability. In contrast, Phikon and PhikonV2 exhibited markedly inferior results across all lesion classes and folds, with drastic performance drops for both MLP ($\Delta$ of –61.19% and –61.85%, respectively) and SVM ($\Delta$ of –61.69% and –64.44%, respectively). This sharp decline highlights a potential domain misalignment: although Phikon models are powerful foundation models trained on very large and heterogeneous histopathology datasets, their embeddings may not encode the finer-grained structural cues required for glomerular lesion discrimination. Another possible factor is that Phikon's large-scale pre-training objective, optimized for broader tissue and slide-level tasks, may overlook the highly localized and subtle morphological features critical in nephropathology.

Taken together, these findings emphasize that not all pathology foundation models are equally transferable to glomerular lesion analysis. While UNI and UNI2 demonstrate stable and accurate performance aligned with task-specific needs, Phikon and PhikonV2 reveal the limitations of using embeddings from models not sufficiently tuned to the microanatomical level of glomeruli. This suggests that selecting, or even adapting, foundation models must carefully consider the granularity and domain-specific requirements of the target pathology.

Figure 3 plots the mean F1-score for each lesion class across varying cross-validation folds (k), comparing the baseline (Fig. 3a) with the two leading FM+classifier configurations: UNI+SVM (Fig. 3b) and UNI2+SVM (Fig. 3c). The trends confirm that embeddings derived from both UNI and UNI2 yield stable classification performance as $k$ increases, in contrast to the baseline. For a detailed breakdown, Fig. 4 presents per-class F1-scores for our highest-performing models in 5-fold cross-validation.

Table III average mean GPU memory utilization alongside average inference latency (ms) per sample. All benchmarks considered both feature extraction and classification time over an NVIDIA RTX 4090 (24 GB VRAM), using identical batch sizes and input resolutions. Models built on UNI embeddings achieve the lowest latency while limiting memory footprint to mid-range levels.

## V. Discussion and Conclusion Remarks

The central hypothesis of our study is that foundation models can achieve competitive performance in glomerular lesion classification without fine-tuning. Our comparative evaluation supports this claim, showing that SVMs leveraging FM-derived embeddings, particularly from UNI and UNI2, consistently outperform an end-to-end CNN baseline (EfficientNet-B0). These results validate the role of foundation models as powerful feature extractors in nephropathology. However, the substantial variability observed across different FMs highlights that scale and pretraining scope alone do not guarantee transferability. Despite their similar training regimes, only UNI and UNI2 produced embeddings with the robustness and stability required for our proprietary dataset.

Future work should therefore focus on strategies to enhance the resilience of foundation models to clinical heterogene-

ity and out-of-distribution data, such as domain-adaptation techniques, task-aware fine-tuning, or the integration of more diverse and pathology-specific pretraining corpora.

## ACKNOWLEDGMENT

David Lima is sponsored by UFBA under the grant no. 51802. Angelo Duarte is sponsored by FAPESB and UEFS, under the grants PET 0017/2024 and FINAPESQ 115/2024. Luciano Oliveira and Washington LC dos-Santos are sponsored by CNPq under grants 301789/2025-8 and 406141/2023.

## REFERENCES

[1] J. R. Weinstein and S. Anderson, "The aging kidney: Physiological changes," *Advances in Chronic Kidney Disease*, vol. 17, no. 4, pp. 302–307, Jul. 2010.

[2] G. Barros, D. Wanderley, L. Rebouças, W. Santos, A. Duarte, and F. Vidal, "Podnet: Ensemble-based classification of podocytopathy on kidney glomerular images," in *Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, 2022, pp. 405–412.

[3] L. Barisoni, K. J. Lafata, S. M. Hewitt, A. Madabhushi, and U. G. J. Balis, "Digital pathology and computational image analysis in nephropathology," *Nature Reviews. Nephrology*, vol. 16, no. 11, pp. 669–685, Nov. 2020.

[4] G. O. Barros, B. Navarro, A. Duarte, and W. L. C. dos Santos, "Pathospotter-k: A computational tool for the automatic identification of glomerular lesions in histological images of kidneys," *Scientific Reports*, vol. 7, no. 1, p. 46769, Apr 2017.

[5] E. Uchino, K. Suzuki, N. Sato, R. Kojima, Y. Tamada, S. Hiragi, H. Yokoi, N. Yugami, S. Minamiguchi, H. Haga, M. Yanagita, and Y. Okuno, "Classification of glomerular pathological findings using deep learning and nephrologist–ai collective intelligence approach," *International Journal of Medical Informatics*, vol. 141, p. 104231, 2020.

[6] J. Besusparis, M. Morkunas, and A. Laurinavicius, "A spatially guided machine-learning method to classify and quantify glomerular patterns of injury in histology images," *Journal of Imaging*, vol. 9, no. 10, 2023.

[7] R. Yamaguchi, Y. Kawazoe, K. Shimamoto, E. Shinohara, T. Tsukamoto, Y. Shintani-Domoto, H. Nagasu, H. Uozaki, T. Ushiku, M. Nangaku, N. Kashihara, A. Shimizu, M. Nagata, and K. Ohe, "Glomerular classification using convolutional neural networks based on defined annotation criteria and concordance evaluation among clinicians," *Kidney International Reports*, vol. 6, no. 3, pp. 716–726, 2021.

[8] P. Chagas, L. Souza, I. Araújo, N. Aldeman, A. Duarte, M. Angelo, W. L.C. dos Santos, and L. Oliveira, "Classification of glomerular hypercellularity using convolutional features and support vector machine," *Artificial Intelligence in Medicine*, vol. 103, p. 101808, 2020.

[9] Y. Nan, F. Li, P. Tang, G. Zhang, C. Zeng, G. Xie, Z. Liu, and G. Yang, "Automatic fine-grained glomerular lesion recognition in kidney pathology," *Pattern Recognition*, vol. 127, p. 108648, 2022.

[10] S. Zhang and D. Metaxas, "On the challenges and perspectives of foundation models for medical image analysis," *Medical Image Analysis*, vol. 91, p. 102996, 2024.

[11] R. Chen, T. Ding, M. Lu, D. F. Williamson, G. Jaume, A. Zhang, B. Chen, A. Zhang, D. Shao, M. Shaban, M. Williams, L. Oldenburg, L. Weishaupt, J. Wang, A. Vaidya, L. Le, G. Gerber, S. Sahai, W. Williams, and F. Mahmood, "Towards a general-purpose foundation model for computational pathology," *Nature Medicine*, vol. 30, no. 3, pp. 850–862, 2024.

[12] A. Filiot, R. Ghermi, A. Olivier, P. Jacob, L. Fidon, A. Camara, A. Mac Kain, C. Saillard, and J.-B. Schiratti, "Scaling Self-Supervised Learning for Histopathology with Masked Image Modeling," 2023.

[13] A. Filiot, P. Jacob, A. M. Kain, and C. Saillard, "Phikon-v2, a large and public feature extractor for biomarker prediction," 2024. [Online]. Available: https://arxiv.org/abs/2409.09173

[14] X. Wang, S. Yang, J. Zhang, M. Wang, J. Zhang, W. Yang, J. Huang, and X. Han, "Transformer-based unsupervised contrastive learning for histopathological image classification," *Medical Image Analysis*, vol. 81, p. 102559, 2022.

[15] A. Kolesnikov, A. Dosovitskiy, D. Weissenborn, H. G., J. Uszkoreit, L. Beyer, M. Minderer, M. Dehghani, N. Houlsby, S. Gelly, T. Unterthiner, and X. Zhai, "An image is worth 16x16 words: Transformers for image recognition at scale," 2021.

[16] J. Breen, K. Allen, K. Zucker, L. Godson, N. M. Orsi, and N. Ravikumar, "A comprehensive evaluation of histopathology foundation models for ovarian cancer subtype classification," *NPJ Precision Oncology*, vol. 9, p. 33, Jan. 2025.

[17] E. D. de Jong, E. Marcus, and J. Teuwen, "Current pathology foundation models are unrobust to medical center differences," 2025.

[18] X. Li, J. Merkow, N. C. F. Codella, A. Santamaria-Pang, N. Sangani, A. Ersoy, C. Burt, J. W. Garrett, R. J. Bruce, J. D. Warner, T. Bradshaw, I. Tarapov, M. P. Lungren, and A. B. McMillan, "From embeddings to accuracy: Comparing foundation models for radiographic classification," 2025. [Online]. Available: https://arxiv.org/abs/2505.10823

[19] K. Enda, Y. Oda, Z. ichi Tanei, K. Satoh, H. Motegi, T. Shunsuke, S. Yamaguchi, T. Ogawa, W. Lei, M. Tsuda, and S. Tanaka, "Transfer learning strategies for pathological foundation models: A systematic evaluation in brain tumor classification," 2025. [Online]. Available: https://arxiv.org/abs/2501.11014

[20] D. Li, G. Wan, X. Wu, X. Wu, A. J. Nirmal, C. G. Lian, P. K. Sorger, Y. R. Semenov, and C. Zhao, "A survey on computational pathology foundation models: Datasets, adaptation strategies, and evaluation tasks," 2025.

[21] E. Vorontsov, A. Bozkurt, A. Casson, G. Shaikovski, M. Zelechowski, K. Severson, E. Zimmermann, J. Hall, N. Tenenholtz, N. Fusi, E. Yang, P. Mathieu, A. van Eck, D. Lee, J. Viret, E. Robert, Y. K. Wang, J. D. Kunz, M. C. H. Lee, J. H. Bernhard, R. A. Godrich, G. Oakley, E. Millar, M. Hanna, H. Wen, J. A. Retamero, W. A. Moye, R. Yousfi, C. Kanan, D. S. Klimstra, B. Rothrock, S. Liu, and T. J. Fuchs, "A foundation model for clinical-grade computational pathology and rare cancers detection," *Nature Medicine*, vol. 30, no. 10, pp. 2924–2935, Oct 2024.

[22] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural Computation*, vol. 1, no. 4, pp. 541–551, 1989.

[23] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[24] M. Caron, H. Touvron, I. Misra, H. Jegou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 9630–9640.

[25] M. Oquab, T. Darcet, T. Moutakanni, H. V. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, R. Howes, P.-Y. Huang, H. Xu, V. Sharma, S.-W. Li, W. Galuba, M. Rabbat, M. Assran, N. Ballas, G. Synnaeve, I. Misra, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski, "Dinov2: Learning robust visual features without supervision," 2023.

[26] J. Zhou, C. Wei, H. Wang, W. Shen, C. Xie, A. Yuille, and T. Kong, "ibot: Image bert pre-training with online tokenizer," 2022. [Online]. Available: https://arxiv.org/abs/2111.07832

[27] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," *arXiv preprint arXiv:2002.05709*, 2020.

[28] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 9726–9735.

[29] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning*, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 18–24 Jul 2021, pp. 8748–8763.

[30] J. Yu, Z. Wang, V. Vasudevan, L. Yeung, M. Seyedhosseini, and Y. Wu, "Coca: Contrastive captioners are image-text foundation models," *Transactions on Machine Learning Research*, vol. Aug 2022, 2022.

[31] M. Bilal, Aadam, M. Raza, Y. Altherwy, A. Alsuhaibani, A. Abduljabbar, F. Almarshad, P. Golding, and N. Rajpoot, "Foundation models in computational pathology: A review of challenges, opportunities, and impact," 2025.