BioMedInformatics

MDPI

Article

Comparative Evaluation of CNN, Vision Transformer, and Swin Transformer Feature Representations with CatBoost for Bethesda Classification of Multi-Cell Cervical Cytology Image Patches

Miguel Angel Valles-Coral [1,*], Ciro Rodriguez [1,*], Carlos Navarro [1], Ulises Román-Concha [1] and Diego Rodriguez [2]

[1] Facultad de Ingeniería de Sistemas e Informática, Universidad Nacional Mayor de San Marcos (UNMSM), Lima 15081, Peru; cnavarrod@unmsm.edu.pe (C.N.); nromanc@unmsm.edu.pe (U.R.-C.)
[2] Facultad de Ciencias de la Salud, Medicina, Universidad Peruana de Ciencias Aplicadas (UPC), Lima 15067, Peru; u202115933@upc.edu.pe
* Correspondence: miguel.valles@unmsm.edu.pe (M.A.V.-C.); crodriguezro@unmsm.edu.pe (C.R.); Tel.: +51-942023414 (M.A.V.-C.); +51-990427136 (C.R.)

Abstract

Artificial intelligence in cervical cytology requires controlled comparisons to distinguish the contribution of visual representations within the classification pipeline. This study evaluated six deep feature representation architectures for Bethesda classification of nucleus-centered multi-cell cervical cytology patches. The CRIC Cervix Collection comprised 400 whole-slide images (WSIs) and 11,530 annotated cellular records across six Bethesda categories. DenseNet121, DINOv2 ViT-B/14, Swin Transformer Tiny, VMamba-T, GFNet-S, and MLP-Mixer-B/16 were evaluated under frozen and partial fine-tuning conditions using CatBoost as a common downstream classifier and WSI-level partitioning to reduce data leakage. In the ten-fold post-selection Dev80 evaluation, DenseNet121 achieved the highest mean Macro-F1 under partial fine-tuning (0.5038 ± 0.0318) and frozen extraction (0.4858 ± 0.0305). However, on the predefined locked Test10 evaluation after Final90 training, partially fine-tuned DINOv2 achieved the strongest overall performance, with a Macro-F1 of 0.4965, Balanced Accuracy of 0.5563, Macro ROC-AUC of 0.9155, and LogLoss of 0.8579. Exploratory Friedman analyses indicated differences among representations, although DenseNet121 and DINOv2 were not statistically distinguished after Holm correction. These findings show that internal stability and locked-test performance may yield different model rankings and support leakage-aware evaluation of heterogeneous visual representations in cervical cytology.

Keywords: Pap smear; deep features; CRIC; hybrid learning; multiclass classification

Check for updates

Academic Editor: James C. L. Chow

Received: 14 July 2026
Revised: 19 August 2026
Accepted: 15 September 2026
Published: 17 September 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Accurate and timely identification of cervical cellular lesions remains a critical component of cervical cancer prevention programs. Cervical cytology, particularly the Papanicolaou test, enables the early recognition of cellular abnormalities that may precede invasive disease and supports clinical decision-making through standardized reporting systems such as the Bethesda System [1]. However, the interpretation of Pap smear samples continues to rely heavily on expert visual assessment, which may be affected by interobserver variability, subjective morphological interpretation, visual fatigue, and differences in professional experience [2]. These limitations may compromise the consistency of cytological

classification, especially in settings where the volume of samples is high and the availability of specialized personnel is limited [3,4].

The clinical impact of cytological interpretation errors extends beyond diagnostic performance indicators [5]. As reported in [6,7], false-positive results may lead to unnecessary complementary examinations, invasive procedures, psychological distress, increased healthcare costs, and avoidable patient burden. Conversely, Macios and Nowakowski [8] reported that false-negative results may delay the identification of clinically relevant lesions and reduce the opportunity for timely follow-up and treatment. These concerns are particularly relevant in regional public health laboratories, where cytological screening programs must operate under substantial workload pressure [9]. In the San Martín region of Peru, the Reference Regional Public Health Laboratory processes approximately 10,000 Papanicolaou tests per year, a volume that can exceed the operational capacity of specialized staff and affect diagnostic timeliness, reproducibility, and quality assurance. This context motivates the development of computational decision-support strategies capable of assisting cytological assessment through objective, scalable, and reproducible image analysis, as explored in [3,10,11].

Artificial intelligence has increasingly been explored as a complementary strategy for improving cervical cytology analysis [12–14]. Studies such as [15–18] have applied convolutional neural networks, transfer learning, Vision Transformers, ensemble models, and hybrid machine learning schemes to automate the classification of cervical cytology images. A previous systematic review of artificial intelligence in cervical cytology showed that deep learning-based approaches have grown substantially, with particular attention to model architectures, datasets, and performance metrics [19]. Nevertheless, this body of evidence also reveals persistent challenges related to dataset representativeness [20,21], inconsistent evaluation protocols [8], limited external validation [22], heterogeneous reporting practices, and the frequent use of controlled datasets that do not fully reflect real cytological workflows.

A further issue in this field is the need for precise clinical and computational terminology. Studies such as [23–25] frame automated cytology classification as "cervical cancer diagnosis", although Pap smear analysis primarily supports the detection and classification of cytological abnormalities that require subsequent clinical confirmation. Therefore, artificial intelligence models in this domain should be understood as tools for Bethesda-based cytological lesion classification rather than as direct substitutes for histopathological cancer diagnosis [26]. This distinction is essential to avoid overclaiming the clinical scope of automated models and to ensure their responsible integration into diagnostic support workflows [19].

From a methodological perspective, published approaches such as [27–29] compare complete pipelines in which feature extractors, classifiers, preprocessing strategies, data partitions, and evaluation metrics vary simultaneously. This makes it difficult to determine whether performance differences are mainly attributable to the visual representation paradigm, the downstream classifier, or other experimental factors. Moreover, several datasets commonly used in cervical cytology consist of isolated single-cell images or highly controlled cropped samples, whereas real cytological interpretation involves the examination of slide regions containing multiple cells, overlapping structures, staining variability, artifacts, inflammatory elements, and local morphological context [30–32]. Datasets such as the one described in [33] enable the extraction of nucleus-centered multi-cell cytology image patches as an intermediate and methodologically relevant unit of analysis, preserving local cellular context while still allowing controlled computational experimentation.

Previous work has demonstrated the feasibility of leakage-free hybrid deep learning strategies for multiclass cervical cytology classification by combining DenseNet-based

feature extraction with different machine learning classifiers [10]. However, that line of work focused on classifier comparison using a single deep visual extractor. More broadly, existing comparative studies often evaluate architectures within a particular family, modify several stages of the pipeline simultaneously, or maintain a single visual representation while varying the downstream classifier. Consequently, the comparative behavior of feature representations derived from substantially different computer vision paradigms remains insufficiently explored when the classification stage is kept constant. This gap is relevant because convolutional networks, self-supervised Vision Transformers, hierarchical shifted-window Transformers, state-space models, frequency-domain networks, and MLP-based architectures encode visual information through fundamentally different mechanisms, which may influence their ability to represent nuclear morphology, cytoplasmic patterns, cellular arrangement, texture, and local contextual information in cervical cytology images.

To address this gap, the present study adopts a controlled hybrid representation–classification framework in which the downstream classifier is held constant while the visual representation paradigm is systematically varied. DenseNet121 was used as the convolutional representative [34], DINOv2 ViT-B/14 as the self-supervised Vision Transformer representative [35], Swin Transformer Tiny as the hierarchical shifted-window Transformer representative [36,37], VMamba-T as the visual state-space representative [38], GFNet-S as the frequency-domain representative [39], and MLP-Mixer-B/16 as the token- and channel-mixing representative [40]. Each backbone was independently coupled with CatBoost, producing six heterogeneous hybrid representation–classification pipelines. Each pipeline was then evaluated under both frozen feature extraction and partial fine-tuning, resulting in twelve directly comparable experimental configurations. CatBoost was maintained as the common downstream classifier across all configurations, reducing variation attributable to the decision stage and enabling a more controlled assessment of how different representation paradigms behave under equivalent classification conditions [41]. All pipelines were evaluated using the same WSI-level partitions, preprocessing framework, hyperparameter-selection procedure, grouped internal analysis, locked independent test partition, and complementary performance metrics.

The main contributions of this study are fourfold. First, it introduces a controlled hybrid comparative framework that systematically couples a common CatBoost classifier with six heterogeneous visual representation paradigms spanning convolutional, self-supervised attention-based, hierarchical Transformer, state-space, frequency-domain, and MLP-based mechanisms. Second, it evaluates each representation under two directly comparable adaptation scenarios—frozen feature extraction and partial fine-tuning—yielding twelve hybrid pipelines and enabling the interaction between representation paradigm and adaptation strategy to be examined under a consistent downstream classification stage. Third, it applies a leakage-aware WSI-level validation strategy suitable for nucleus-centered multi-cell cervical cytology patches, reducing the risk of inflated performance estimates caused by related cellular samples from the same WSI appearing across different partitions. Fourth, it frames the task as Bethesda-based cytological lesion classification, avoiding overstatement of direct cancer diagnosis and aligning the computational task with the clinical role of Pap smear screening.

2. Related Work

2.1. CNN and Transfer Learning Approaches in Cervical Cytology Classification

Convolutional neural networks and transfer learning have been widely applied to cervical cytology image classification, particularly because pretrained models enable the reuse of visual representations learned from large-scale image datasets when annotated cytological data are comparatively limited. Tran et al. [18] explored transfer learning

for multiclass cervical cytology classification, demonstrating the feasibility of adapting pretrained deep architectures to discriminate among different cellular categories. More recently, Kaur et al. [34] conducted a broader comparison of pretrained CNN architectures using the Herlev and SIPaKMeD datasets. Their results showed that the best-performing architecture varied according to the dataset and class configuration, with different models leading under different classification settings. This variability indicates that reported CNN superiority should be interpreted in relation to the experimental setting rather than as an architecture-independent property.

Similar variability can be observed across other comparative studies. Tan et al. [42] evaluated 13 pretrained CNN models for seven-class Pap smear classification on the Herlev dataset and identified DenseNet-201 as the strongest model in their experimental setting. In contrast, Bhatt et al. [43] investigated transfer learning combined with progressive resizing for cervical cell classification using Pap smear whole-slide image data, introducing a different training strategy and image-level formulation. Although these studies reinforce the effectiveness of pretrained CNNs, their results are not directly interchangeable because the datasets, number of classes, image units, preprocessing procedures, and training strategies differ across experiments.

From a representational perspective, CNNs provide strong inductive biases for learning local morphological and textural characteristics, including nuclear shape, chromatin patterns, cytoplasmic appearance, and staining-related information. However, the existing literature frequently evaluates CNN architectures as part of complete classification pipelines in which several experimental components vary simultaneously. Consequently, differences in reported performance may reflect not only the visual representation produced by the backbone, but also differences in dataset composition, fine-tuning strategy, preprocessing, class formulation, downstream classification, and validation design. CNN-based representations therefore constitute an important reference for cervical cytology analysis, while a controlled comparison with alternative visual representation mechanisms is needed to determine how the representation stage itself behaves when the remaining experimental conditions are kept as consistent as possible.

2.2. Transformer-Based and Attention-Oriented Models

Vision Transformers have gained increasing attention in cervical cytology because self-attention mechanisms enable the modeling of broader spatial relationships beyond the predominantly local patterns captured by conventional convolutional operations. Darwish et al. [3] proposed a Vision Transformer approach for cervical precancerous classification using shifted patch tokenization to improve the learning of discriminative visual patterns. Similarly, Maurya et al. [44] developed VisionCervix, combining Vision Transformer and CNN representations for Pap smear image classification. Their ViT-CNN ensemble achieved higher classification performance than an alternative CNN-LSTM strategy, although at a greater computational cost. This comparison illustrates that incorporating global-context modeling can improve discriminative performance, but the resulting gains may also depend on architectural fusion and computational complexity rather than on the Transformer representation alone.

Hierarchical Transformer architectures have also been investigated to balance local and global representation. Alohali et al. [36] proposed Swin-GA-RF, combining Swin Transformer feature extraction with genetic algorithm-based feature selection and Random Forest classification on SIPaKMeD. In a broader setting, Khan et al. [45] introduced CervixFormer, a multi-scale Swin Transformer framework evaluated on both WSI- and patch-level cervical cytology data, including CRIC Cervix, Mendeley LBC, SIPaKMeD, and Cervix93. Beyond using a hierarchical representation, CervixFormer incorporated synthetic

data generation, multi-scale ensemble learning, feature fusion, and stain-normalization analysis. Its multi-dataset evaluation provides useful evidence regarding the applicability of Swin-based representations across heterogeneous cytology sources, but also illustrates the difficulty of attributing performance improvements to a single architectural component when several methodological mechanisms are introduced simultaneously.

Taken together, these studies show that Transformer-based approaches can represent contextual relationships that are relevant to cervical cytology, but they also reveal substantial heterogeneity in experimental design. Existing studies differ in image unit, dataset composition, feature fusion, augmentation, feature selection, downstream classifier, and validation strategy. Moreover, many approaches evaluate Transformers as components of optimized end-to-end or hybrid pipelines rather than as independently controlled feature representations. Consequently, the reported superiority of a Transformer-based method cannot necessarily be attributed to its representation mechanism alone. This motivates comparisons in which alternative visual representations are evaluated under a more consistent downstream classification and validation framework.

2.3. State-Space, Frequency-Domain, and MLP-Based Visual Representations

Beyond convolutional and attention-based architectures, recent computer vision research has introduced alternative mechanisms for modeling spatial relationships and long-range dependencies. State-space models, frequency-domain networks, and MLP-based architectures approach visual representation without relying exclusively on conventional convolution or self-attention. These alternatives are relevant for medical imaging because they provide different strategies for integrating local and global information while potentially reducing some of the computational constraints associated with attention-based models.

Within state-space approaches, Rodriguez et al. [38] evaluated VMamba-Tiny alongside CNN and Swin Transformer backbones for breast ultrasound image classification and segmentation, illustrating the applicability of selective state-space modeling to medical images. Frequency-domain representations have been explored through GFNet, introduced by Rao et al. [39], which models long-range spatial interactions using learnable global filters in the Fourier domain rather than self-attention. In turn, Ahmed and Nandi [40] evaluated pretrained MLP-Mixer variants, including MLP-Mixer-B/16, for mammographic classification and reported competitive performance across several breast imaging datasets. Although these studies were conducted outside cervical cytology, they demonstrate that state-space, frequency-domain, and token-mixing mechanisms can provide viable alternatives to conventional CNN and Transformer representations.

However, these architectural families remain much less studied in cervical cytology than CNN- and Transformer-based approaches, particularly under controlled comparisons in which the downstream classifier and validation protocol are kept consistent. Existing evidence is also heterogeneous in terms of imaging modality, dataset scale, task formulation, and training strategy, limiting direct conclusions about the relative quality of their feature representations. Their inclusion in a common experimental framework therefore broadens the representational comparison and allows their behavior to be examined under the same classification and evaluation conditions.

2.4. Hybrid Deep Feature Extraction and Machine Learning Classifiers

Hybrid approaches combining deep feature extractors with conventional machine learning classifiers have become increasingly relevant in cervical cytology classification. These methods separate visual representation from the final decision stage: a deep architecture generates high-dimensional embeddings, while classifiers such as SVM, Random Forest, XGBoost, or CatBoost perform the subsequent classification. This separation pro-

vides greater flexibility than fully end-to-end schemes and facilitates the analysis of how representation and classification components contribute to the overall pipeline.

Valles-Coral et al. [10] proposed a leakage-free hybrid framework for multiclass classification of cervical cells in digital Papanicolaou images, combining DenseNet-based feature extraction with different machine learning classifiers. Their study emphasized methodological control and statistical evaluation while reducing the risk of inflated performance estimates caused by data leakage. However, because the visual extractor remained fixed while the classifiers varied, the comparison primarily characterized the influence of the decision stage rather than differences among deep visual representations.

Sahoo et al. [28] followed a different hybrid strategy by combining several pretrained models with complementary information fusion in a multistage classification framework. In contrast to the controlled extractor used by Valles-Coral et al. [10], this design integrates information from multiple components to improve predictive robustness. Although such fusion strategies can enhance classification performance, changing several stages simultaneously makes it more difficult to determine which component is responsible for the observed improvement.

Taken together, these studies illustrate two common directions in hybrid cervical cytology modeling: comparing classifiers over a shared deep representation and combining multiple representations or decision stages. Both approaches provide evidence for the usefulness of hybrid pipelines, but neither directly examines how different visual representation mechanisms behave when the downstream classifier and evaluation framework are held consistent. This distinction motivates a more controlled comparison of deep feature representations under a common classification scheme.

2.5. Cervical Cytology Datasets and Image Units

The design and interpretation of cervical cytology classification studies are strongly influenced by the dataset and image unit used for model development. Zhang et al. [27] introduced DCCL as a benchmark for cervical cytology analysis, contributing to more structured evaluation and comparison of computational methods. Nevertheless, results obtained across different datasets are not necessarily directly comparable because they may differ in acquisition conditions, class definitions, image resolution, staining characteristics, and the level at which cytological information is represented.

A substantial portion of the cervical cytology literature relies on isolated cells or highly controlled cropped images. These units facilitate standardized classification experiments but reduce part of the local context encountered in routine cytological examination. In practice, relevant regions may contain neighboring or overlapping cells, inflammatory elements, artifacts, staining variability, and heterogeneous morphological patterns. Therefore, performance obtained from isolated-cell datasets should be interpreted cautiously when considering methods intended to analyze image regions containing richer cytological context.

The CRIC Cervix Collection described by Rezende et al. [33] provides conventional Pap smear cytology images together with cellular annotations that enable the extraction of nucleus-centered multi-cell patches while preserving surrounding cellular information. These patches constitute an intermediate unit of analysis between isolated-cell crops and complete WSI assessment, allowing for controlled experimentation without removing all local context. This characteristic is particularly relevant when comparing different visual representation mechanisms, since their ability to encode morphology, texture, spatial relationships, and contextual information may vary according to the amount and organization of visual information retained within each patch.

2.6. Critical Synthesis and Methodological Gap

Overall, the reviewed literature shows that cervical cytology classification has progressed through increasingly diverse visual representation strategies, ranging from CNN-based transfer learning and attention-oriented Transformers to hybrid deep feature extraction schemes. More recent state-space, frequency-domain, and MLP-based architectures further expand the available mechanisms for encoding visual information. However, the reported results are difficult to compare directly because studies differ substantially in dataset composition, image unit, number and definition of classes, preprocessing, fine-tuning strategy, feature selection, classifier, and validation design. Consequently, differences in reported performance cannot generally be attributed to the visual representation mechanism alone.

A second limitation concerns how comparative experiments are formulated. Many studies modify several components of the pipeline simultaneously, whereas others maintain a single feature extractor and compare different downstream classifiers. Both strategies provide useful evidence, but they leave less explored the behavior of substantially different visual representations when the classification stage is kept consistent. This issue becomes particularly relevant for multi-cell cervical cytology patches, where local morphology and surrounding cellular context may be encoded differently by convolutional, attention-based, state-space, frequency-domain, and MLP-based architectures.

Accordingly, the unresolved question is not whether deep learning can support cervical cytology classification, but how different visual representation mechanisms behave when evaluated under a common downstream classifier and a consistent, leakage-aware experimental framework. The present study addresses this gap by comparing six deep visual backbones under frozen and partial fine-tuning conditions, while maintaining CatBoost as the downstream classifier and applying WSI-level grouped validation and an independent locked test partition. This design does not eliminate the intrinsic architectural differences among the models, but it reduces variation attributable to the classification stage and enables a more controlled assessment of their representations for Bethesda classification of nucleus-centered multi-cell cervical cytology patches.

3. Materials and Methods

This study followed a quantitative comparative experimental design to evaluate six deep feature representation pipelines for Bethesda classification of multi-cell cervical cytology image patches. The comparative framework was broadened through the incorporation of three additional contemporary architectures, extending the range of visual representation mechanisms assessed under the same experimental conditions. A common protocol was applied for data partitioning, image preprocessing, feature extraction, CatBoost (version 1.2.10) classification, and performance assessment, while each backbone was examined under frozen and partial fine-tuning conditions.

3.1. Dataset Description and Experimental Partition

3.1.1. CRIC Dataset

The CRIC Cervix Collection dataset was used, which contains digital cervical cytology images obtained from Papanicolaou samples [33]. The study considered 400 whole-slide images (WSIs) and an initial set of 11,534 cellular annotations organized into six Bethesda categories: NILM, ASC-US, ASC-H, LSIL, HSIL, and SCC [46]. Each record included the source image, the coordinates of the annotated nucleus, and its diagnostic label. The annotated nucleus defined the class of the corresponding image patch, while other cells within the patch were retained as visual context.

3.1.2. Data Integrity Audit

Before data partitioning, the availability and readability of the images and the validity of the nuclear coordinates were verified. A deterministic identifier (cell_uid) was generated from the image filename and nuclear coordinates to detect repeated or conflicting annotations. Four exact duplicates were removed, and no contradictory labels were identified, resulting in 11,530 valid records: 6777 NILM, 606 ASC-US, 924 ASC-H, 1360 LSIL, 1702 HSIL, and 161 SCC.

The audit also identified 64 pairs of annotated nuclei separated by less than 16 pixels. These records were retained because proximity alone did not indicate duplication. Since the available metadata did not include a patient- or case-level identifier, the WSI was used as the highest grouping unit throughout the experimental design.

3.1.3. WSI-Level Multilabel Experimental Partition

The dataset was partitioned at the WSI level using image_filename as the grouping identifier. Therefore, all cellular records from the same WSI were assigned to a single subset, preventing related samples from appearing across the development and independent test partitions [10]. Because each WSI could contain more than one Bethesda category, the partitioning procedure considered the presence of all classes within each image.

The 400 WSIs were divided into four mutually exclusive subsets: FT10, with 40 WSIs and 1442 records for backbone adaptation; Dev70, with 280 WSIs and 8194 records for CatBoost training; HPO10, with 40 WSIs and 869 records for hyperparameter selection; and Test10, with 40 WSIs and 1025 records for independent evaluation. The partitions were generated using mixed-integer linear programming while maintaining the predefined sizes and class coverage. Test10 remained blocked and was not loaded during model development.

For partial fine-tuning, FT10 was further divided at the WSI level into FT-train, containing 32 WSIs and 1250 records, and FT-validation, containing 8 WSIs and 192 records. This internal split was used to train the adaptable backbone layers and select the corresponding checkpoint before feature extraction. The complete experimental partitioning scheme is summarized in Figure 1.

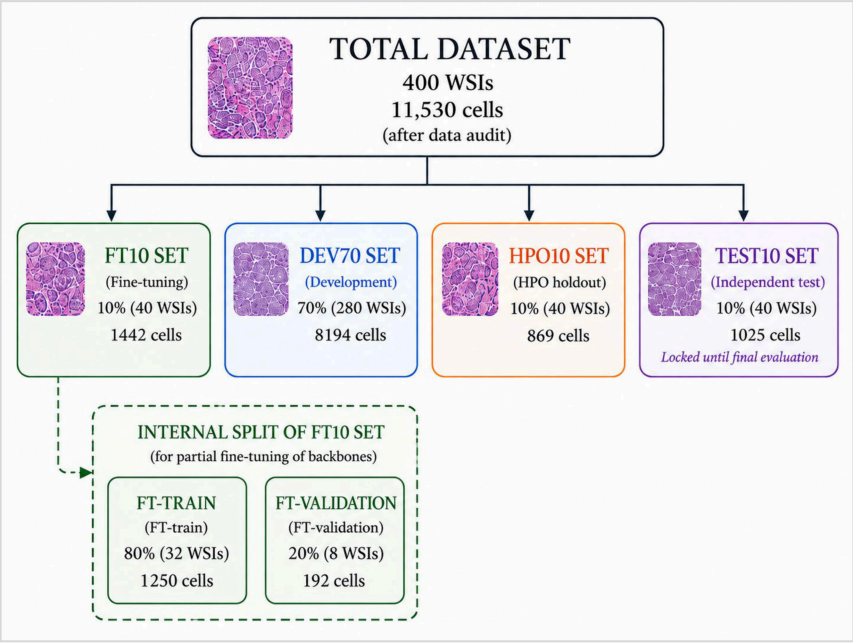


Figure 1. WSI-level experimental partition of the CRIC dataset.

3.2. Image Preprocessing and Data Augmentation

3.2.1. Image Loading and Nucleus-Centered Cropping

Each source image was loaded and converted to RGB. Using the nucleus_x and nucleus_y coordinates provided in the dataset [33], a 160 × 160-pixel patch centered on the annotated nucleus was extracted. When the crop extended beyond the image boundaries, constant zero-padding was applied to preserve both the patch size and the position of the nucleus within the patch. Samples with invalid or out-of-range coordinates had previously been excluded during the data audit. Figure 2a illustrates both the normal cropping case and the near-border case requiring zero-padding.

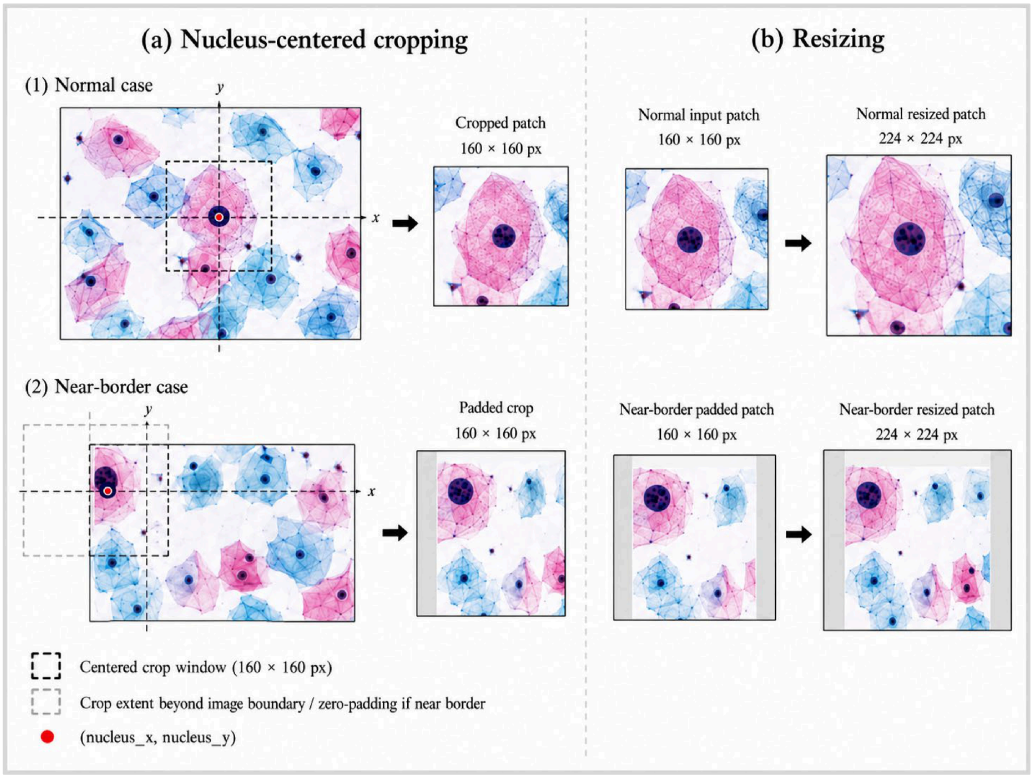


Figure 2. Preparation of nucleus-centered cytology patches through cropping, zero-padding, and resizing.

3.2.2. Image Resizing and Normalization

The extracted patches were resized from 160 × 160 to 224 × 224 pixels and converted into tensors before being processed by the deep feature extractors [47]. Bilinear interpolation and ImageNet normalization—mean [0.485, 0.456, 0.406] and standard deviation [0.229, 0.224, 0.225]—were used for DenseNet121, DINOv2 ViT-B/14, Swin-T, VMamba-T, and GFNet-S. For MLP-Mixer-B/16, bicubic interpolation and a mean and standard deviation of [0.5, 0.5, 0.5] were applied according to its pretrained configuration. The complete preparation procedure for nucleus-centered patches is summarized in Figure 2.

3.2.3. Data Augmentation

During partial fine-tuning, data augmentation was applied only to FT-train. The procedure included random resized cropping with a scale range of 0.85–1.00 and an aspect ratio of 0.90–1.10, horizontal flipping with a probability of 0.50, vertical flipping with a probability of 0.20, rotations of up to 20°, and moderate adjustments to brightness, contrast, saturation, and hue. The same augmentation policy was applied to all Bethesda categories. FT-validation and the datasets used for embedding extraction received only deterministic resizing, tensor conversion, and the normalization corresponding to each pretrained model.

3.3. Deep Feature Representation Models

Six pretrained backbones were used to extract deep representations from the cytology patches: DenseNet121 [48,49], DINOv2 ViT-B/14 [35], Swin Transformer Tiny [37], VMamba-T [38,50], GFNet-S [39], and MLP-Mixer-B/16 [40]. These architectures were selected because they process visual information through different mechanisms. DenseNet121 uses convolutional operations; DINOv2 employs global self-attention; Swin-T applies hierarchical attention through shifted windows; VMamba-T uses visual state-space modeling; GFNet-S incorporates global filtering in the frequency domain; and MLP-Mixer-B/16 combines information through token- and channel-mixing operations.

The inclusion of these models allowed the cytology patches to be represented through complementary approaches, ranging from local morphological processing to broader spatial and global image relationships. For each backbone, the original classification head was removed, and the native output preceding the classification layer was retained as a fixed-length embedding. These representations were subsequently used as input to an independent CatBoost classifier [41].

All models received patches with the same input resolution and were evaluated using the same experimental partitions and downstream classification procedure. Their original pretraining configurations and native embedding dimensions were preserved, as summarized in Table 1.

Table 1. Main configuration of the deep feature representation models.

Model	Representation Paradigm	Pretraining	Embedding Dimension
DenseNet121	Convolutional processing	Supervised ImageNet-1K	1024
DINOv2 ViT-B/14	Global self-attention	DINOv2 self-supervised pretraining	768
Swin-T	Hierarchical shifted-window attention	Supervised ImageNet-1K	768
VMamba-T	Visual state-space modeling	Supervised ImageNet-1K	768
GFNet-S	Global frequency-domain filtering	Supervised ImageNet-1K	384
MLP-Mixer-B/16	Token and channel mixing	ImageNet-21K followed by ImageNet-1K	768

The native embedding dimensions were retained during feature extraction. The frozen and partial fine-tuning conditions applied to the backbones are described in Section 3.4.

3.4. Frozen and Partial Fine-Tuning Protocols

3.4.1. Backbone Conditions

Each backbone was evaluated under two conditions: frozen and partial fine-tuning. In both cases, the models received the preprocessed nucleus-centered patches of 224 × 224 pixels described in Section 3.2. Under the frozen condition, all pretrained backbone parameters remained unchanged, and each model was used directly in evaluation mode to generate its native embeddings.

For partial fine-tuning, all backbone parameters were initially frozen, after which only the final architecture-specific modules were enabled for training. The trainable components were denseblock4 and norm5 for DenseNet121; Transformer blocks 11–12 and the final normalization layer for DINOv2 ViT-B/14; stage 4 and the final normalization layer for Swin-T and VMamba-T; Global Filter blocks 18–19 and the final normalization layer for GFNet-S; and Mixer blocks 11–12 and the final normalization layer for MLP-Mixer-B/16. All remaining modules were kept frozen and in evaluation mode during adaptation.

Each backbone constituted an independent pipeline and generated its own embedding representation, which was subsequently supplied to a separate CatBoost classifier. Em-

beddings from different backbones were not combined. The frozen and partial fine-tuning conditions are summarized in Figure 3.

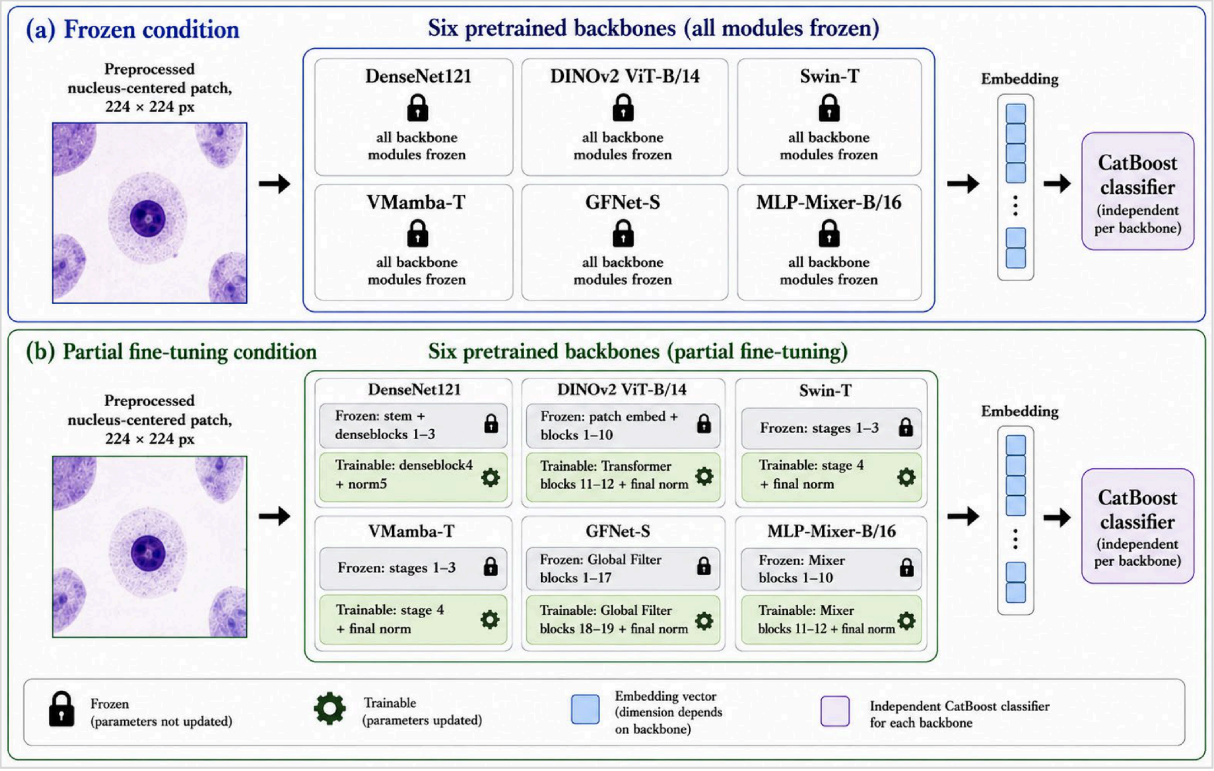


Figure 3. Frozen and partial fine-tuning conditions applied to the six deep feature extractors. Each backbone generated an independent native embedding that was subsequently evaluated using its corresponding CatBoost classifier.

3.4.2. Partial Fine-Tuning Procedure

Partial fine-tuning was performed using FT-train, while FT-validation was reserved for checkpoint selection [51]. A temporary classification head composed of a dropout layer with a probability of 0.20 and a linear layer with six outputs was attached to each backbone. Only this temporary head and the architecture-specific modules described in Section 3.4.1 were optimized.

Training was conducted for a maximum of 16 epochs using the AdamW optimizer implemented in PyTorch (version 2.11.0), with a learning rate of 10×10^{-4} and a weight decay of 10×10^{-4}. A physical batch size of 8 and gradient accumulation over two steps produced an effective batch size of 16, while a batch size of 16 was used for validation. Class imbalance was addressed through weighted cross-entropy, with the class weights calculated exclusively from FT-train. Weighted sampling was not applied, and data augmentation was restricted to FT-train as described in Section 3.2.3.

The checkpoint with the highest Macro-F1 on FT-validation was retained. When two epochs achieved the same Macro-F1, the lower validation loss was used as the secondary selection criterion. Early stopping was applied after six consecutive epochs without improvement. After checkpoint selection, the temporary classification head was removed, and the selected backbone was placed in evaluation mode to generate the embeddings used in the subsequent CatBoost classification stage.

3.5. Embedding Extraction and Reproducibility Controls

After establishing the backbone state for each experimental condition, embeddings were extracted separately for every data subset. The backbones were placed in evaluation

mode, gradient computation was disabled, and the data loaders were configured without sample shuffling or removal. Each observation was required to contain its image path, image_filename, Bethesda label, nuclear coordinates, and unique cell_uid. The indices returned during extraction were used to preserve the original order of the observations.

The resulting embedding matrices and encoded labels were stored together with metadata containing the cell_uid, source image, diagnostic label, and nuclear coordinates. Before storage, the procedure verified that the number and order of embeddings, labels, and metadata were consistent and that no duplicated cell_uid values were present. These controls maintained the correspondence between each numerical representation and its original cellular annotation.

To prevent the reuse of incompatible embeddings, each stored set was associated with fingerprints of the backbone state, experimental subset, preprocessing configuration, and image content. Previously generated embeddings were reused only when these fingerprints, the backbone condition, and the number of observations matched the current execution; otherwise, they were extracted again. Reproducibility was additionally controlled using a fixed seed of 42 for Python (version 3.13.15), NumPy (version 2.1.3), PyTorch (version 2.11.0), CUDA, and the data loaders, together with deterministic PyTorch and cuDNN settings.

3.6. CatBoost Classification and Hyperparameter Optimization

CatBoost was trained independently for each combination of backbone and experimental condition. The classifier was configured for six-class prediction using the MultiClass objective, Bayesian bootstrap, automatic class balancing through auto_class_weights = Balanced, and a fixed random seed of 42. The iteration-based overfitting detector was enabled with a patience of 40 iterations.

Hyperparameter optimization was performed using the Tree-structured Parzen Estimator implemented in Optuna (version 4.9.0), with 20 completed trials for each pipeline. Every trial was trained on Dev70 and evaluated on the fixed HPO10 subset, while Test10 remained inaccessible. Cross-validation was not applied during this stage because HPO10 was specifically reserved for model configuration selection.

The search included 200–500 iterations in increments of 50, tree depths from 4 to 6, learning rates from 0.02 to 0.15, L2 leaf regularization from 1.0 to 8.0, random strength from 10×10^{-3} to 2.0, and bagging temperatures from 0.0 to 3.0. PCA was evaluated as an optional preprocessing component, using either 64 or 128 components. When enabled, PCA was fitted exclusively on Dev70 and then applied to HPO10 to prevent information leakage [52,53].

Macro-F1 on HPO10 was used as the primary optimization criterion. Accuracy, macro-precision, macro-recall, one-versus-rest ROC-AUC, and multiclass LogLoss were also recorded for descriptive analysis. Ties in Macro-F1 were resolved successively using lower LogLoss, shallower tree depth, lower PCA dimensionality, and fewer fitted iterations. The selected hyperparameters, representation configuration, and number of iterations were then fixed for the post-selection evaluation and final model training described in Section 3.7.

3.7. Post-Selection Internal Analysis and Final Classifier Training

After completing hyperparameter optimization, the selected configuration of each backbone-condition pipeline was subjected to an internal stability analysis on Dev80, obtained by combining Dev70 and HPO10. This partition comprised 320 WSI and 9063 cellular records. Ten-fold grouped cross-validation with WSI-level multilabel stratification was applied, using image_filename as the grouping variable. Consequently, all cells originating from the same WSI remained within a single fold.

The previously selected CatBoost configuration and number of iterations remained fixed throughout the 10 folds, and no early stopping or additional hyperparameter selection was performed. When PCA was included in the selected pipeline, it was fitted using only the training portion of each fold and subsequently applied to the corresponding validation portion. Predictions generated for the validation folds were combined to obtain out-of-fold metrics and to examine the consistency of the selected configuration across different WSI groups.

Because this analysis was conducted after configuration selection, its results were used only to characterize internal stability and were not employed to modify or reselect the pipelines. Final generalization performance was determined exclusively through the locked Test10 evaluation described in Section 3.8.

For final training, the CatBoost classifier corresponding to each pipeline was fitted on Final90, formed by combining FT10, Dev70, and HPO10, resulting in 360 WSI and 10,505 cellular records. The previously established backbone state remained unchanged during this stage. CatBoost was trained using the fixed representation configuration, hyperparameters, and number of iterations, without an additional validation subset or early stopping. When selected, PCA was fitted on Final90 before classifier training. Test10 remained excluded until the final evaluation stage.

3.8. Locked Test Evaluation and Performance Assessment

Final generalization performance was assessed on Test10, comprising 40 WSI and 1025 cellular records. This partition remained inaccessible throughout model development and was opened only after the backbone states, representation configurations, CatBoost hyperparameters, and training iterations had been fixed. Test10 did not contribute to pipeline selection or adjustment.

For each backbone-condition pipeline, the primary evaluation was performed using the CatBoost classifier trained on Final90. A classifier trained on Dev80 was additionally assessed as a prespecified sensitivity analysis. Its results were used only to examine the consistency of the observed performance and did not influence the selected configurations.

Performance was quantified using Accuracy, macro-precision, macro-recall, Macro-F1, multiclass one-versus-rest ROC-AUC, and LogLoss [54]. Precision, recall, and F1-score were also calculated for each Bethesda category. Macro-F1 was treated as the primary summary metric, while confusion matrices and class-wise ROC and precision–recall curves were used to examine prediction patterns and class-specific behavior.

3.9. Uncertainty Estimation and Exploratory Statistical Analysis

Uncertainty in Test10 performance was estimated through grouped bootstrap resampling, using image_filename as the sampling unit to preserve the dependence among cellular patches originating from the same WSI. A total of 2000 bootstrap samples were generated with replacement, and 95% percentile bootstrap confidence intervals were obtained from the 2.5th and 97.5th percentiles for Accuracy, macro-precision, macro-recall (Balanced Accuracy), Macro-F1, multiclass one-versus-rest ROC-AUC, and LogLoss. For ROC-AUC, confidence intervals were calculated using only valid bootstrap replicates containing all six Bethesda categories.

Exploratory comparisons among the six backbone–CatBoost pipelines were performed separately for the frozen and partial fine-tuning conditions using the paired Macro-F1 values obtained from the 10 shared internal Dev80 folds. Before analysis, the equality of WSI assignments across pipelines was verified. Descriptive statistics included the mean, standard deviation, median, interquartile range, minimum, maximum, and average rank.

The global comparison was conducted using the Friedman test, with Kendall's W reported as an effect-size measure. Pairwise contrasts were examined using two-sided Wilcoxon signed-rank tests with Pratt's treatment of zero differences. The resulting 15 p-values were adjusted using the Holm procedure [54]. Mean and median differences, together with rank-biserial correlations, were reported to describe the direction and magnitude of the observed contrasts. Statistical significance was established at $\alpha = 0.05$.

Because these comparisons were performed after pipeline configuration and the cross-validation folds shared training observations, the statistical findings were interpreted as exploratory evidence of internal consistency rather than confirmatory proof of superiority. Conclusions regarding generalization were based on the locked Test10 results and their WSI-grouped bootstrap confidence intervals.

4. Results

4.1. Hyperparameter Optimization and Selected Pipeline Configurations

Hyperparameter optimization produced different CatBoost configurations for the frozen and partial fine-tuning representations. Among the partial pipelines, DenseNet121 + CatBoost obtained the highest Macro-F1 on HPO10, reaching 0.5297, followed by MLP-Mixer-B/16 + CatBoost with 0.4922 and DINOv2 ViT-B/14 + CatBoost with 0.4742. Under the frozen condition, DenseNet121 + CatBoost also achieved the highest value, with a Macro-F1 of 0.5172.

PCA was retained in four of the twelve selected pipelines: partial DINOv2 ViT-B/14 with 128 components, partial VMamba-T and MLP-Mixer-B/16 with 64 components, and frozen Swin-T with 64 components. The remaining pipelines preserved the native dimensionality of their embeddings. The main CatBoost configurations selected for each pipeline are summarized in Table 2.

Table 2. Main CatBoost configurations selected on HPO10 for the frozen and partial fine-tuning pipelines.

Condition	Backbone	PCA	Fixed Iterations	Depth	HPO10 Macro-F1
Partial	DenseNet121	-	271	6	0.5297
	DINOv2 ViT-B/14	128	110	5	0.4742
	Swin-T	-	78	6	0.4650
	VMamba-T	64	456	5	0.4112
	GFNet-S	-	227	6	0.4005
	MLP-Mixer-B/16	64	155	4	0.4922
Frozen	DenseNet121	-	102	6	0.5172
	DINOv2 ViT-B/14	-	313	6	0.4222
	Swin-T	64	128	5	0.4183
	VMamba-T	-	82	6	0.3944
	GFNet-S	-	215	6	0.3686
	MLP-Mixer-B/16	-	263	5	0.4329

Note. PCA values indicate the number of retained components; "-" indicates that PCA was not applied. Fixed iterations correspond to the number retained after early stopping and subsequently used for the post-selection internal analysis and final CatBoost training.

Descriptively, the selected partial fine-tuning configuration achieved a higher HPO10 Macro-F1 than its frozen counterpart for all six backbones. The largest difference was observed for MLP-Mixer-B/16, whereas DenseNet121 showed the closest values between conditions. These results were used to establish the configurations subsequently evaluated on Dev80 and were not interpreted as independent estimates of generalization performance.

4.2. Post-Selection Internal Stability on Dev80

Across the 10 WSI-grouped folds on Dev80, DenseNet121 + CatBoost achieved the highest mean Macro-F1 under both experimental conditions, as summarized in Table 3. The partial fine-tuning pipeline reached a Macro-F1 of 0.5038 ± 0.0318, together with the highest mean Accuracy and macro-averaged one-versus-rest ROC-AUC and the lowest LogLoss. Its frozen counterpart obtained a Macro-F1 of 0.4858 ± 0.0305.

Table 3. Post-selection internal performance on Dev80 across 10 WSI-grouped folds. Values are presented as mean ± standard deviation.

Panel A. Partial Fine-Tuning				
Backbone	**Accuracy**	**Macro-F1**	**Macro ROC-AUC(OvR)**	**LogLoss**
DenseNet121	0.6933 ± 0.0506	0.5038 ± 0.0318	0.9126 ± 0.0195	0.8215 ± 0.1126
DINOv2 ViT-B/14	0.6875 ± 0.0403	0.4817 ± 0.0402	0.9064 ± 0.0148	0.8616 ± 0.0726
Swin-T	0.6399 ± 0.0567	0.4623 ± 0.0396	0.8960 ± 0.0204	0.9062 ± 0.1083
VMamba-T	0.6113 ± 0.0596	0.4385 ± 0.0358	0.8744 ± 0.0168	1.0185 ± 0.0994
GFNet-S	0.6032 ± 0.0603	0.4370 ± 0.0341	0.8654 ± 0.0189	1.1361 ± 0.0809
MLP-Mixer-B/16	0.6556 ± 0.0599	0.4662 ± 0.0498	0.8989 ± 0.0228	0.8916 ± 0.1062
Panel B. Frozen				
Backbone	**Accuracy**	**Macro-F1**	**Macro ROC-AUC (OvR)**	**LogLoss**
DenseNet121	0.6745 ± 0.0521	0.4858 ± 0.0305	0.8978 ± 0.0190	0.8917 ± 0.1048
DINOv2 ViT-B/14	0.6385 ± 0.0529	0.4566 ± 0.0221	0.8836 ± 0.0189	0.9869 ± 0.0860
Swin-T	0.6204 ± 0.0597	0.4357 ± 0.0426	0.8680 ± 0.0205	1.0030 ± 0.1283
VMamba-T	0.6059 ± 0.0500	0.4157 ± 0.0317	0.8624 ± 0.0180	1.0440 ± 0.0895
GFNet-S	0.5795 ± 0.0754	0.4137 ± 0.0435	0.8458 ± 0.0218	1.2067 ± 0.0684
MLP-Mixer-B/16	0.6194 ± 0.0577	0.4245 ± 0.0330	0.8576 ± 0.0195	1.0537 ± 0.0894

Under partial fine-tuning, DINOv2 ViT-B/14 achieved the second-highest mean Macro-F1, with 0.4817 ± 0.0402, followed by MLP-Mixer-B/16 and Swin-T, with 0.4662 ± 0.0498 and 0.4623 ± 0.0396, respectively. VMamba-T and GFNet-S showed similar mean Macro-F1 values of 0.4385 ± 0.0358 and 0.4370 ± 0.0341. In the frozen condition, DINOv2 ViT-B/14 ranked second according to mean Macro-F1, followed by Swin-T, MLP-Mixer-B/16, VMamba-T, and GFNet-S.

For all six backbones, the partial fine-tuning representation achieved a higher mean Macro-F1 than its frozen counterpart. The absolute differences ranged from 0.0180 for DenseNet121 to 0.0417 for MLP-Mixer-B/16. These findings characterize the internal stability of the fixed configurations across the shared WSI-level folds. They were not used for further pipeline selection and should not be interpreted as independent estimates of generalization performance.

4.3. Locked Test Evaluation and Overall Performance

The twelve pipelines were evaluated on the locked Test10 partition after final CatBoost training on Final90, as summarized in Table 4. Given the class imbalance of the dataset, Macro-F1 was considered the primary metric for model comparison, complemented by Balanced Accuracy, Macro ROC-AUC, and LogLoss. Under partial fine-tuning, DINOv2 + CatBoost achieved the strongest overall performance, with a Macro-F1 of 0.4965, Balanced Accuracy of 0.5563, Macro ROC-AUC of 0.9155, and the lowest LogLoss of 0.8579. MLP-Mixer-B/16 ranked second according to Macro-F1 (0.4691), followed by Swin-T (0.4322), DenseNet121 (0.4230), VMamba-T (0.4004), and GFNet-S (0.3794).

Table 4. Overall performance on the locked Test10 partition after final CatBoost training on Final90.

Mode	Backbone + CatBoost	Balanced Accuracy	Macro-F1	Macro ROC-AUC	LogLoss
Partial	DINOv2	0.5563	0.4965	0.9155	0.8579
	MLP-Mixer-B/16	0.5353	0.4691	0.8795	1.0164
	Swin-T	0.4905	0.4322	0.8811	1.0475
	DenseNet121	0.4776	0.4230	0.8856	1.0204
	VMamba-T	0.4468	0.4004	0.8644	1.0618
	GFNet-S	0.4350	0.3794	0.8530	1.1776
Frozen	MLP-Mixer-B/16	0.5013	0.4649	0.8632	1.0713
	Swin-T	0.4958	0.4482	0.8623	1.0232
	DINOv2	0.4952	0.4355	0.8814	1.0110
	DenseNet121	0.4646	0.4172	0.8842	0.9511
	VMamba-T	0.4030	0.3915	0.8478	1.0735
	GFNet-S	0.4408	0.3859	0.8421	1.2372

Within the frozen condition, MLP-Mixer-B/16 achieved the highest Macro-F1 (0.4649), followed by Swin-T (0.4482) and DINOv2 (0.4355). DenseNet121 reached a Macro-F1 of 0.4172 but showed the highest Macro ROC-AUC (0.8842) and lowest LogLoss (0.9511) among the frozen pipelines. VMamba-T and GFNet-S obtained Macro-F1 values of 0.3915 and 0.3859, respectively. The resulting ranking therefore differed from the internal Dev80 evaluation, in which DenseNet121 had obtained the highest mean Macro-F1.

The effect of partial fine-tuning on Test10 was model-dependent rather than uniformly beneficial. Compared with the corresponding frozen representations, Macro-F1 increased for DINOv2, MLP-Mixer-B/16, DenseNet121, and VMamba-T, whereas Swin-T and GFNet-S showed slightly lower values after partial fine-tuning. The largest increase was observed for DINOv2, from 0.4355 to 0.4965. This behavior contrasts with the internal Dev80 analysis, where all six backbones showed higher mean Macro-F1 under partial fine-tuning, highlighting the importance of the locked test evaluation for assessing generalization.

A sensitivity analysis was additionally performed by training the final CatBoost classifier on Dev80 instead of Final90 while retaining the same locked Test10 partition. Figure 4 illustrates the resulting changes in Macro-F1 for the partial and frozen conditions, while the corresponding numerical values are reported in Table 5. Under partial fine-tuning, MLP-Mixer-B/16 and DINOv2 achieved the highest Dev80-sensitivity Macro-F1 values, at 0.4861 and 0.4847, respectively. Under the frozen condition, Swin-T and DINOv2 reached 0.5034 and 0.5008. The largest differences relative to Final90 were observed for frozen DINOv2 (+0.0653) and Swin-T (+0.0553).

Table 5. Sensitivity of Test10 Macro-F1 to the final CatBoost training partition.

Mode	Backbone + CatBoost	Final90	Dev80 sensitivity	Δ
Partial	DINOv2	0.4965	0.4847	−0.0118
	MLP-Mixer-B/16	0.4691	0.4861	+0.0169
	Swin-T	0.4322	0.4184	−0.0137
	DenseNet121	0.4230	0.4670	+0.0440
	VMamba-T	0.4004	0.4187	+0.0183
	GFNet-S	0.3794	0.4178	+0.0383

Table 5. *Cont.*

Mode	Backbone + CatBoost	Final90	Dev80 sensitivity	Δ
Frozen	MLP-Mixer-B/16	0.4649	0.4556	−0.0093
	Swin-T	0.4482	0.5034	+0.0553
	DINOv2	0.4355	0.5008	+0.0653
	DenseNet121	0.4172	0.4674	+0.0502
	VMamba-T	0.3915	0.4331	+0.0417
	GFNet-S	0.3859	0.4077	+0.0218

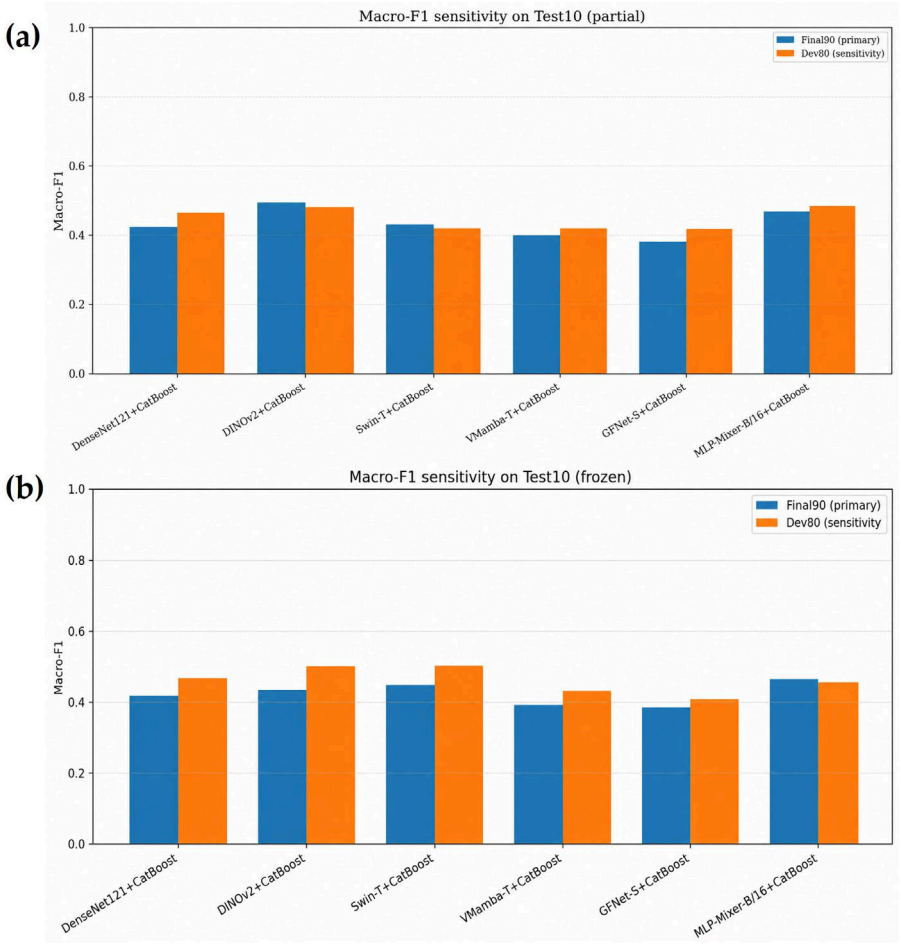


Figure 4. Sensitivity of Macro-F1 on the locked Test10 partition to the final CatBoost training partition under (**a**) partial fine-tuning and (**b**) frozen feature extraction. Final90 represents the predefined primary evaluation, whereas Dev80 represents the sensitivity analysis.

Overall, DINOv2 with partial fine-tuning provided the strongest result under the predefined primary Final90 evaluation, leading the twelve pipelines in Macro-F1, Balanced Accuracy, Macro ROC-AUC, and LogLoss. Nevertheless, the sensitivity analysis showed that the relative ordering of the representations was not completely invariant to the data used for final classifier training. Accordingly, Final90 was retained as the primary evaluation for the subsequent class-wise analysis, whereas Dev80 was considered only as a sensitivity assessment.

4.4. Class-Wise Performance Analysis

Class-wise performance on the locked Test10 partition showed substantial variation across the six Bethesda categories. Table 6 reports the F1-score obtained by each pipeline

after final CatBoost training on Final90. The class distribution in Test10 comprised 49 ASC-H, 52 ASC-US, 171 HSIL, 109 LSIL, 625 NILM, and 19 SCC samples. Among the partial fine-tuning pipelines, DINOv2 achieved the highest F1-score for ASC-US (0.2644), LSIL (0.6186), and NILM (0.8466), while MLP-Mixer-B/16 obtained the highest partial F1 for ASC-H (0.2871) and HSIL (0.5000), and SCC performance was also highest with MLP-Mixer-B/16 (0.5854).

Table 6. Class-wise F1-score on the locked Test10 partition using the primary Final90 evaluation.

Panel A. Partial Fine-Tuning						
Bethesda Class	**DenseNet121**	**DINOv2**	**Swin-T**	**VMamba-T**	**GFNet-S**	**MLP-Mixer-B/16**
ASC-H	0.2100	0.2513	0.2424	0.1927	0.2121	0.2871
ASC-US	0.2581	0.2644	0.2092	0.2326	0.2053	0.2245
HSIL	0.3228	0.4268	0.4476	0.3701	0.4408	0.5000
LSIL	0.5231	0.6186	0.5000	0.4224	0.4123	0.4240
NILM	0.7996	0.8466	0.7366	0.7730	0.7031	0.7938
SCC	0.4242	0.5714	0.4571	0.4118	0.3030	0.5854
Panel B. Frozen						
Bethesda Class	**DenseNet121**	**DINOv2**	**Swin-T**	**VMamba-T**	**GFNet-S**	**MLP-Mixer-B/16**
ASC-H	0.2264	0.3086	0.2047	0.1547	0.2471	0.2098
ASC-US	0.2414	0.2373	0.1932	0.1875	0.1974	0.2513
HSIL	0.4765	0.5895	0.5804	0.4240	0.5145	0.5733
LSIL	0.4769	0.3985	0.3228	0.3305	0.3559	0.3891
NILM	0.8029	0.7458	0.7655	0.7855	0.7146	0.7713
SCC	0.2791	0.3333	0.6222	0.4667	0.2857	0.5946

Across both conditions, NILM was the most consistently recognized category, with F1-scores ranging from 0.7031 to 0.8466. In contrast, ASC-H and ASC-US showed the lowest F1 values for most pipelines, ranging from 0.1547 to 0.3086 and from 0.1875 to 0.2644, respectively. Performance for HSIL and LSIL varied more substantially according to the representation and adaptation strategy. SCC also showed considerable variability, with F1-scores between 0.2791 and 0.6222; however, this category contained only 19 Test10 samples, so its class-specific estimates should be interpreted cautiously.

For the primary best-performing pipeline, DINOv2 with partial fine-tuning, the class-wise results further illustrate differences between precision and recall. NILM achieved a precision of 0.9178 and recall of 0.7856, while LSIL reached 0.5748 and 0.6697, respectively. SCC showed a precision of 0.5217 and recall of 0.6316. In contrast, ASC-H and ASC-US presented recalls of 0.5102 and 0.4423 but substantially lower precisions of 0.1667 and 0.1885. HSIL exhibited the opposite pattern, with high precision (0.7500) but lower recall (0.2982). These differences explain why overall accuracy alone would provide an incomplete description of performance across the Bethesda categories.

The normalized confusion matrix of DINOv2 with partial fine-tuning is shown in Figure 5. The diagonal values correspond to class-specific recall and reached 0.7856 for NILM, 0.6697 for LSIL, and 0.6316 for SCC, while HSIL showed a lower correct-classification proportion of 0.2982. ASC-H and ASC-US reached 0.5102 and 0.4423, respectively. The matrix therefore confirms that the strongest overall pipeline did not perform uniformly across categories and that errors remained concentrated in the more difficult abnormal classes.

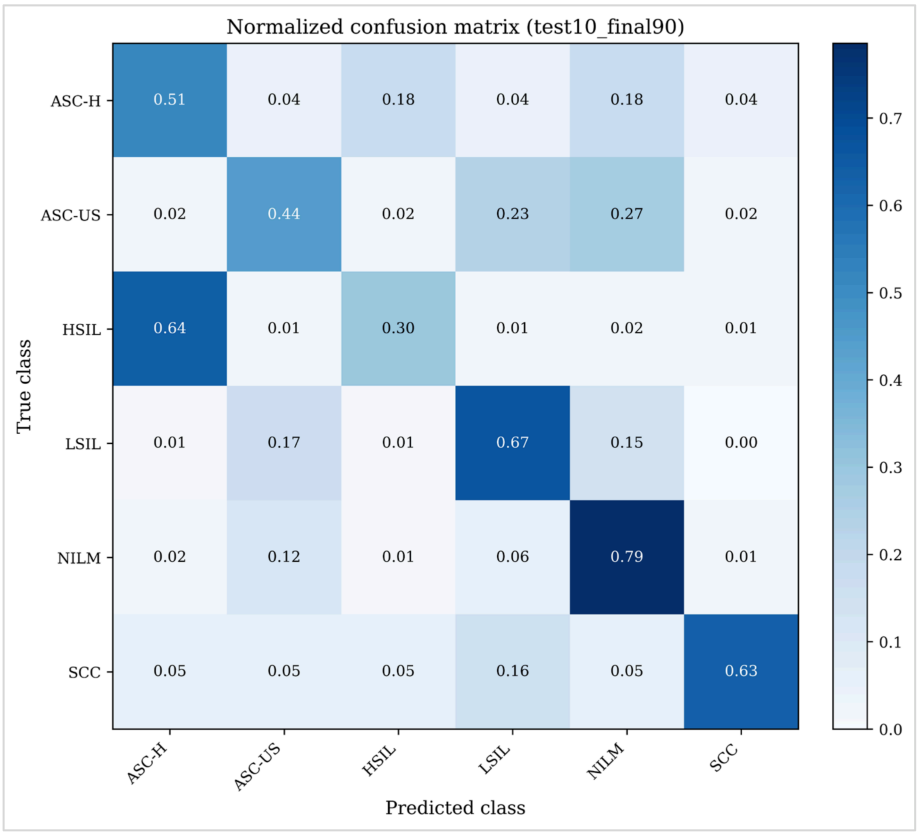


Figure 5. Normalized confusion matrix of DINOv2 + CatBoost with partial fine-tuning on the locked Test10 partition using the primary Final90 evaluation. Rows represent true Bethesda categories and columns represent predicted categories.

4.5. Uncertainty Estimates and Exploratory Statistical Results

Uncertainty in the locked Test10 evaluation was characterized using 2000 WSI-grouped bootstrap replicates. Confidence intervals were calculated for the global performance metrics of both the Final90 primary evaluation and the Dev80 sensitivity analysis. Because Final90 constituted the prespecified primary evaluation, the results presented below focus on this setting. No pairwise inferential testing was performed on Test10.

Given that Macro-F1 was the primary summary metric, Table 7 reports its observed Test10 values and corresponding 95% bootstrap confidence intervals for all six representations under both experimental conditions. DINOv2 achieved the highest point estimate under partial fine-tuning, whereas MLP-Mixer-B/16 obtained the highest value under frozen extraction. However, the intervals overlapped substantially across several pipelines, showing that the relative ordering observed from the point estimates was accompanied by considerable uncertainty.

Table 7. Macro-F1 estimates and WSI-grouped bootstrap 95% confidence intervals on the locked Test10 partition after Final90 training.

Visual Representation	Partial Fine-Tuning Macro-F1 [95% CI]	Frozen Extraction Macro-F1 [95% CI]
DenseNet121	0.4230 [0.3097–0.4821]	0.4172 [0.3181–0.4585]
DINOv2 ViT-B/14	0.4965 [0.3775–0.5371]	0.4355 [0.3126–0.4774]
Swin-T	0.4322 [0.2990–0.4755]	0.4482 [0.3118–0.5068]
VMamba-T	0.4004 [0.3085–0.5026]	0.3915 [0.2674–0.5393]
GFNet-S	0.3794 [0.2703–0.4352]	0.3859 [0.2784–0.4291]
MLP-Mixer-B/16	0.4691 [0.3110–0.5283]	0.4649 [0.3214–0.5783]

Figure 6 complements these numerical estimates by showing the WSI-grouped bootstrap distributions. The partial fine-tuning condition preserved the higher observed performance of DINOv2, while the frozen condition showed a more competitive distribution for MLP-Mixer-B/16 and Swin-T. The width and overlap of the distributions also illustrate the variability associated with evaluating the pipelines on the 40 WSIs contained in Test10.

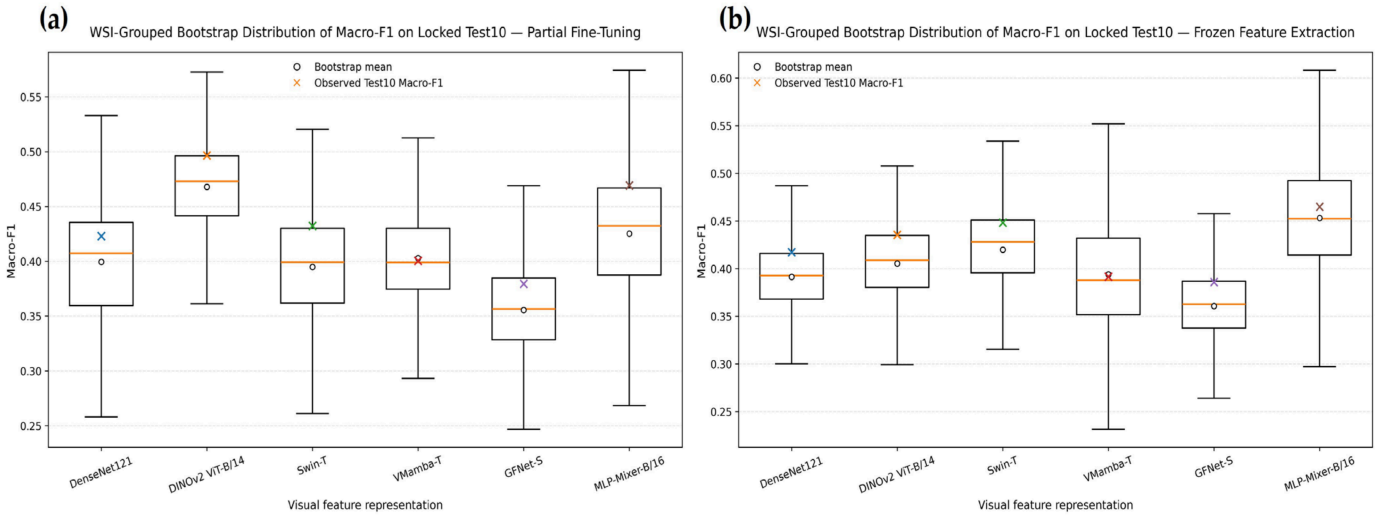


Figure 6. WSI-grouped bootstrap distributions of Macro-F1 on the locked Test10 partition after Final90 training under (**a**) partial fine-tuning and (**b**) frozen feature extraction. The boxplots summarize the distributions obtained from 2000 WSI-level bootstrap replicates. Open circles indicate the bootstrap mean, whereas crosses indicate the observed Test10 Macro-F1.

Separately, the post-selection Dev80 folds were used to explore the internal consistency of the six representations. Friedman testing indicated differences in Macro-F1 among the pipelines under both frozen and partial fine-tuning conditions, with similar Kendall's W values in the two analyses (Table 8). These results indicate that the model ordering was not random across folds, although the analysis remained exploratory because the folds were not independent experimental replicates.

Table 8. Exploratory Friedman analysis of Macro-F1 across the ten shared post-selection Dev80 folds.

Condition	Friedman χ^2	p	Kendall's W
Frozen	28.8571	0.000025	0.5771
Partial fine-tuning	29.0857	0.000022	0.5817

DenseNet121 achieved the best average rank in both conditions, followed by DINOv2. Figure 7 shows that this ordering was accompanied by fold-to-fold variability, particularly among the intermediate-performing representations. These distributions describe the internal stability of the fixed configurations and were not used for further model selection.

The Holm-adjusted pairwise comparisons showed that the clearest internal differences were concentrated between the higher-ranked representations and several lower-performing pipelines (Table 9). Importantly, DenseNet121 and DINOv2 were not statistically distinguished after correction in either condition, despite DenseNet121 obtaining the better average rank. This result supports a cautious interpretation of the internal ranking and is consistent with the change in ordering subsequently observed on the locked Test10 partition.

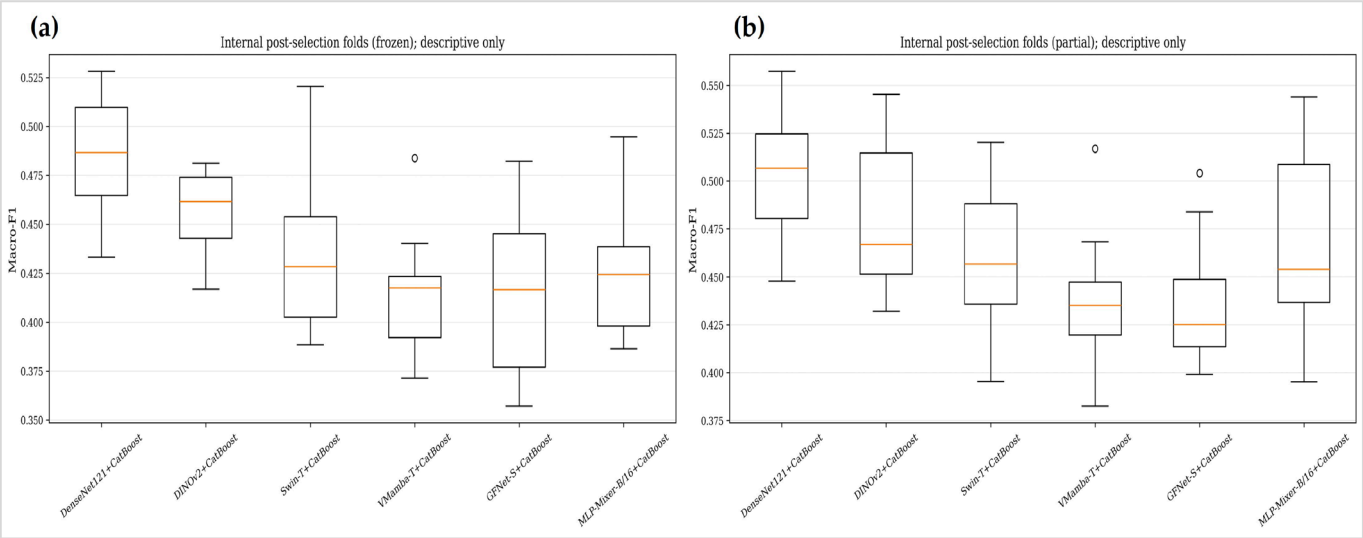


Figure 7. Distribution of Macro-F1 across the ten shared post-selection Dev80 folds for the six visual representations under (**a**) frozen feature extraction and (**b**) partial fine-tuning. The boxplots are descriptive because the folds share training observations. Open circles indicate observations beyond the boxplot whiskers.

Table 9. Pairwise comparisons remaining significant after Holm correction. Positive rank-biserial values favor the first model.

Condition	Comparison	Holm-Adjusted p	Rank-Biserial
Frozen	DenseNet121 vs. Swin-T	0.0293	1.0000
	DenseNet121 vs. VMamba-T	0.0293	1.0000
	DenseNet121 vs. GFNet-S	0.0293	1.0000
	DenseNet121 vs. MLP-Mixer-B/16	0.0293	1.0000
	DINOv2 vs. GFNet-S	0.0430	0.9636
Partial	DenseNet121 vs. VMamba-T	0.0293	1.0000
	DenseNet121 vs. GFNet-S	0.0469	0.9636
	DenseNet121 vs. MLP-Mixer-B/16	0.0469	0.9636
	DINOv2 vs. VMamba-T	0.0469	0.9636
	DINOv2 vs. GFNet-S	0.0293	1.0000
	Swin-T vs. VMamba-T	0.0293	1.0000

Overall, the bootstrap analysis quantified the uncertainty surrounding the locked Test10 results, whereas the Dev80 comparisons characterized the internal consistency of the selected pipelines. Accordingly, the exploratory statistical findings were not treated as confirmatory evidence of architectural superiority, and conclusions regarding generalization remained based on the predefined locked Test10 evaluation and its WSI-grouped confidence intervals.

5. Discussion

The results reveal that the relative performance of the visual representations depended on the evaluation stage rather than being dominated by a single architecture throughout the experiment. DenseNet121 + CatBoost showed the strongest internal consistency in Dev80, obtaining the highest mean Macro-F1 under both frozen and partial fine-tuning conditions. However, on the locked Test10 partition, DINOv2 + CatBoost with partial fine-tuning achieved the strongest primary Final90 result, with a Macro-F1 of 0.4965, Balanced Accuracy of 0.5563, Macro ROC-AUC of 0.9155, and the lowest LogLoss of 0.8579. The

sensitivity analysis also showed changes in model ordering when CatBoost was trained on Dev80 instead of Final90. Therefore, the findings indicate that internal ranking and locked-test generalization were not completely equivalent.

The strong internal behavior of DenseNet121 is consistent with previous studies highlighting the usefulness of convolutional architectures and transfer learning in cervical cytology [18,34]. CNN representations are well suited to encode local morphological and textural patterns associated with nuclear appearance, cytoplasmic characteristics, and staining variability. Nevertheless, the superior locked-test performance of partially fine-tuned DINOv2 indicates that convolutional representations were not uniformly advantageous. The common CatBoost classifier reduced variation attributable to the decision stage, allowing for a more controlled comparison of the feature spaces generated by the different backbones, although differences intrinsic to their pretraining and architectural mechanisms remained.

Regarding Transformer-based representations, the results support their applicability but also show that their behavior depends on the adaptation strategy. Previous studies have emphasized the ability of Vision Transformers and Swin Transformers to capture broader contextual relationships in cervical images [3,36], whereas hybrid approaches often combine representation learning with feature selection, fusion, or multiple classifiers [28]. In the present experiment, partial fine-tuning substantially improved DINOv2 on Test10, whereas Swin-T showed a slightly lower Macro-F1 than its frozen counterpart. Moreover, partial fine-tuning was not uniformly beneficial across all six backbones. This suggests that adaptation to cervical cytology is architecture-dependent rather than an automatic consequence of updating pretrained parameters.

The broader comparison also showed that useful representations were not restricted to CNN and Transformer families. MLP-Mixer-B/16 was particularly competitive, ranking second in the primary partial fine-tuning evaluation and first among the frozen pipelines according to Macro-F1, whereas VMamba-T and GFNet-S generally showed lower performance. These findings extend previous hybrid evidence such as [10], where deep feature extraction was combined with machine learning classifiers under a leakage-aware design. Unlike approaches centered on varying the downstream classifier, the present study maintained CatBoost across all six backbones and used WSI-level grouping, thereby reducing classification-stage variability and preventing patches derived from the same slide from being distributed across training and evaluation partitions.

Class-wise results further showed that overall performance did not translate into uniform recognition across Bethesda categories. NILM was the most consistently classified category, while ASC-H and ASC-US remained the most difficult for most pipelines. For the best primary pipeline, DINOv2 with partial fine-tuning, LSIL and NILM achieved comparatively strong F1-scores, whereas HSIL showed high precision but substantially lower recall. SCC also presented relatively high performance for some models, but its Test10 representation was limited to 19 samples, making its class-specific estimates more variable. These results reinforce the importance of Macro-F1 and class-wise analysis in addition to aggregate discrimination metrics, particularly in an imbalanced multiclass cytology setting.

The exploratory Friedman analyses indicated differences among the six representations under both frozen and partial conditions, with similar Kendall's W values of approximately 0.58. However, the pairwise comparisons did not statistically distinguish DenseNet121 from DINOv2 after Holm correction in either condition. On the locked Test10 partition, the WSI-grouped bootstrap confidence intervals showed substantial overlap across several pipelines, indicating uncertainty in the relative ordering suggested by the point estimates. Together with the change in ranking observed between Dev80 and Test10, these findings

caution against interpreting the internal ordering as definitive evidence of architectural superiority. Overall, the findings support DINOv2 with partial fine-tuning as the strongest pipeline in the predefined locked-test evaluation, while DenseNet121 demonstrated the greatest internal post-selection consistency.

Within this controlled experimental framework, some limitations should nevertheless be considered when interpreting the findings. The evaluation was conducted using the CRIC Cervix Collection, and some Bethesda categories were less represented than others, which may increase the uncertainty of class-specific estimates for minority categories. Although the WSI-level partitioning and predefined locked Test10 evaluation provided a rigorous assessment within the available data, external validation on independent cervical cytology datasets would be valuable to assess the transferability of the observed behavior across different acquisition, staining, and institutional conditions.

6. Conclusions

This study demonstrates the importance of evaluating deep visual representations under a common and leakage-aware experimental framework for Bethesda classification of multi-cell cervical cytology patches. By maintaining CatBoost as the downstream classifier across six backbones and applying WSI-level partitioning, the comparison reduced variability attributable to the classification stage while preventing patches from the same slide from being distributed across training and evaluation subsets. DenseNet121 showed the greatest internal post-selection consistency in Dev80, whereas DINOv2 with partial fine-tuning achieved the highest observed performance in the predefined locked Test10 evaluation, reaching a Macro-F1 of 0.4965 (95% CI: 0.3775–0.5371), Balanced Accuracy of 0.5563, Macro ROC-AUC of 0.9155, and LogLoss of 0.8579.

The results also showed that partial fine-tuning was not uniformly beneficial across architectures. DINOv2 exhibited the largest improvement on the locked test set, while other backbones showed smaller gains or slight decreases relative to their frozen representations. MLP-Mixer-B/16 also demonstrated competitive performance, whereas VMamba-T and GFNet-S generally obtained lower results under the evaluated conditions. At the class level, NILM was the most consistently recognized category, while ASC-H and ASC-US remained more challenging. These findings indicate that performance depends not only on the architectural family but also on the adaptation strategy and the characteristics of the cytological categories being modeled.

Methodologically, the study provides a controlled comparison of convolutional, attention-based, state-space, frequency-domain, and MLP-based visual representations using a common classifier, grouped internal validation, and an independent locked test partition. The exploratory statistical analysis supported internal differences among representations, although DenseNet121 and DINOv2 were not statistically distinguished after Holm correction. In addition, the WSI-grouped bootstrap confidence intervals on Test10 showed substantial overlap among several pipelines, indicating uncertainty in their relative ordering and supporting a cautious interpretation of point-estimate rankings. Future studies should extend the evaluation to external datasets, increase representation of minority Bethesda categories, explore broader domain-adaptation strategies, and incorporate visual interpretability methods. Overall, DINOv2 with partial fine-tuning provided the highest observed locked-test result, while DenseNet121 showed the most stable internal behavior across post-selection folds.

7. Recommendations

Future studies should expand the evaluation of this comparative framework using datasets with greater variability in acquisition conditions, staining protocols, sample prepa-

ration, and institutional origin. For this purpose, the compatibility between available diagnostic categories and the Bethesda classes used in this study should be considered beforehand, since differences in nomenclature or label grouping may limit direct comparisons across datasets.

It is also recommended to maintain leakage-aware partitioning strategies, preferably at the WSI, patient, or acquisition-source level, when multiple cellular samples originate from the same image or slide. This criterion is essential to avoid artificially inflated performance estimates and to obtain evaluations that more closely reflect real-world generalization.

Finally, future research should explore broader domain-adaptation and fine-tuning strategies across different visual representation architectures and strengthen methods for handling imbalance among Bethesda categories, particularly for underrepresented classes. External validation should also be prioritized to assess the stability of the observed rankings across datasets. In addition, visual interpretability methods should be incorporated to determine whether model predictions rely on clinically relevant patterns such as nuclear morphology, cytoplasmic characteristics, and local cellular organization.

Author Contributions: Conceptualization: M.A.V.-C., C.R. and U.R.-C.; Data Curation: M.A.V.-C., C.N. and U.R.-C.; Methodology: M.A.V.-C., C.R. and C.N.; Investigation: M.A.V.-C., C.R., U.R.-C. and D.R.; Formal Analysis: M.A.V.-C., C.N. and U.R.-C.; Software: M.A.V.-C., C.N. and D.R.; Resources: M.A.V.-C., C.R. and D.R.; Supervision: M.A.V.-C. and C.R.; Validation: M.A.V.-C., C.N. and D.R.; Visualization: M.A.V.-C. and U.R.-C.; Writing—Original Draft: M.A.V.-C. and U.R.-C.; Writing—Review & Editing: M.A.V.-C., C.R., C.N., U.R.-C. and D.R.; Project Administration: M.A.V.-C. and C.R.; Funding Acquisition: C.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by Universidad Nacional Mayor de San Marcos—RR No. 007064-2026-R/UNMSM and Project number C26200111–Project type PCONFIGI and year 2026.

Institutional Review Board Statement: Not applicable. This study used a publicly available and previously anonymized cervical cytology image dataset; therefore, no new human participants were recruited and no additional biological samples were collected.

Informed Consent Statement: Not applicable. The study was conducted using a publicly available dataset with anonymized records and did not involve direct interaction with human participants.

Data Availability Statement: The CRIC Cervix Collection dataset used in this study is publicly available at: https://cricdatabase.com.br/ (accessed on 13 July 2026), as described by Rezende et al. [33].

Acknowledgments: The authors acknowledge Universidad Nacional Mayor de San Marcos for the support provided for the execution of this research under RR No. 007064-2026-R/UNMSM, Project No. C26200111, Project Type PCONFIGI, 2026. The authors also acknowledge the creators and maintainers of the CRIC Cervix Collection for making the cervical cytology image dataset used in this study publicly available. During the preparation of this manuscript, the authors used ChatGPT, GPT-5.5 Thinking version (OpenAI, https://chatgpt.com/), as an auxiliary tool for language revision, text structuring, and improvement of writing clarity. The authors reviewed, edited, and approved all content generated with the support of this tool and take full responsibility for the final content of the publication.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Anonymous. The 1988 Bethesda System for Reporting Cervical/Vaginal Cytological Diagnoses. *JAMA J. Am. Med. Assoc.* **1989**, *262*, 931. [CrossRef]
2. Papanicolaou, G.N.; Traut, H.F. The Diagnostic Value of Vaginal Smears in Carcinoma of the Uterus. *Am. J. Obstet. Gynecol.* **1941**, *42*, 193–206. [CrossRef]

3. Darwish, M.; Altabel, M.Z.; Abiyev, R.H. Enhancing Cervical Pre-Cancerous Classification Using Advanced Vision Transformer. *Diagnostics* **2023**, *13*, 2884. [CrossRef]
4. Wang, P.; Wang, J.; Li, Y.; Li, L.; Zhang, H. Adaptive Pruning of Transfer Learned Deep Convolutional Neural Network for Classification of Cervical Pap Smear Images. *IEEE Access* **2020**, *8*, 50674–50683. [CrossRef]
5. Almonte, M.; Murillo, R.; Sánchez, G.I.; González, P.; Ferrera, A.; Picconi, M.A.; Wiesner, C.; Cruz-Valdez, A.; Lazcano-Ponce, E.; Jerónimo, J.; et al. Multicentric study of cervical cancer screening with human papillomavirus testing and assessment of triage methods in Latin America: The ESTAMPA screening study protocol. *BMJ Open* **2020**, *10*, e035796. [CrossRef]
6. Kumar, R.R.; Kumar, V.A.; Kumar, P.N.S.; Sudhamony, S.; Ravindrakumar, R. Detection and removal of artifacts in cervical cytology images using support vector machine. In Proceedings of the ITME 2011—IEEE International Symposium on IT in Medicine and Education, Guangzhou, China, 9–11 December 2011; pp. 717–721. [CrossRef]
7. Castro-Apodaca, F.; Cortez-Hernández, J.A.; García-Ocaranza, P.E.; Canizalez-Román, A.; Barajas-Olivas, F.; Favela-Heredia, C.E.; Magaña-Ordorica, D.; Magaña-Gómez, J.A.; Peña-García, G.M.; Murillo-Llanes, J. Diagnostic Performance of Cytocolposcopy versus Biopsy for Premalignant and Malignant Lesions in a Women's Hospital Dysplasia Clinic. *AL-Kindy Coll. Med. J.* **2024**, *20*, 176–181. [CrossRef]
8. Macios, A.; Nowakowski, A. False Negative Results in Cervical Cancer Screening—Risks, Reasons and Implications for Clinical Practice and Public Health. *Diagnostics* **2022**, *12*, 1508. [CrossRef]
9. Pinedo, J.C.S.M.; López, C.A.R.; Grández, C.R.; Estrella, C.W.G. Reconocimiento de patrones de imágenes a través de un sistema de visión artificial en MATLAB. *Rev. Cient. Sist. Inform.* **2021**, *1*, 15–26. [CrossRef]
10. Valles-Coral, M.A.; Navarro-Cabrera, J.R.; Pinedo, L.; Cotrina-Linares, J.; Sánchez-Flores, J.; Arévalo-Ramirez, H.; Arévalo-Fasanando, L.; Reátegui-Lozano, N.; Injante, R. Mathematical Modeling and Statistical Evaluation of Hybrid Deep Learning Architectures for Multiclass Classification of Cervical Cells in Digital Papanicolaou Images. *Mathematics* **2026**, *14*, 1139. [CrossRef]
11. Gaona, Y.J.; Malla, D.C.; Crespo, B.V.; Vicuña, M.J.; Neira, V.A.; Dávila, S.; Verhoeven, V. Radiomics Diagnostic Tool Based on Deep Learning for Colposcopy Image Classification. *Diagnostics* **2022**, *12*, 1694. [CrossRef]
12. Karampotsis, E.; Dounias, G.; Jantzen, J. Machine Learning Approaches for Pap-Smear Diagnosis: An Overview. In *Machine Learning Paradigms*; Springer: Cham, Switzerland, 2019; Volume 1. [CrossRef]
13. Harsono, A.B.; Susiarno, H.; Suardi, D.; Owen, L.; Fauzi, H.; Kireina, J.; Wahid, R.A.; Carolina, J.S.; Mantilidewi, K.I.; Hidayat, Y.M. Cervical pre-cancerous lesion detection: Development of smartphone-based VIA application using artificial intelligence. *BMC Res. Notes* **2022**, *15*, 356. [CrossRef]
14. de Souza, P.F.G.; Junior, G.G.d.S.T.; de Macedo, D.D.J. Obstetric Decision-Support System: An informational model for maternal autonomy towards the Agenda 2030 health goals. In *Advanced Notes in Information Science*; Pro-Metrics: Tallinn, Estonia, 2025. [CrossRef]
15. AlMohimeed, A.; Shehata, M.; El-Rashidy, N.; Mostafa, S.; Saleh, H.; Talaat, A.S. ViT-PSO-SVM: Cervical Cancer Predication Based on Integrating Vision Transformer with Particle Swarm Optimization and Support Vector Machine. *Bioengineering* **2024**, *11*, 729. [CrossRef]
16. Murugan, K.; Gomathy, B.; Prasath, N.; Umaamaheshvari, A.; Malathi, D. A Super-Resolution Technique with An Ensemble Deep Learning Based for Cervical Cancer Detection System. In Proceedings of the 2022 IEEE North Karnataka Subsection Flagship International Conference Nkcon 2022, Vijaypur, India, 20–21 November 2022. [CrossRef]
17. Dong, M.; Wang, Y.; Zang, Z.; Todo, Y. BiNext-Cervix: A novel hybrid model combining BiFormer and ConvNext for Pap smear classification. *Appl. Intell.* **2025**, *55*, 144. [CrossRef]
18. Tran, V.-K.; Do, D.-T.; Dinh, X.-L.; Nghiem, C.-C. Transfer Learning for Cervical Cancer Multi-class Classification. In *Lecture Notes in Networks and Systems*; Springer: Cham, Switzerland, 2025; Volume 1205. [CrossRef]
19. Valles-Coral, M.; Pinedo, L.; Rodríguez, C.; Rodríguez, D.; Sánchez-Dávila, K.; Arévalo-Fasanando, L.; Reátegui-Lozano, N. Application of artificial intelligence in cervical cytology: A systematic review of deep learning models, datasets, and reported metrics. *Front. Big Data* **2026**, *8*, 1678863. [CrossRef]
20. Anand, A.; Panicker, J.V. Transformer-based Multiclass Classification of Cervicogram Images for Improved Cervical Cancer Detection. In Proceedings of the 2023 14th International Conference on Computing Communication and Networking Technologies ICCCNT 2023, Delhi, India, 6–8 July 2023. [CrossRef]
21. Wubineh, B.Z.; Rusiecki, A.; Halawa, K. Data Augmentation Techniques to Detect Cervical Cancer Using Deep Learning: A Systematic Review. In *Lecture Notes in Networks and Systems*; Springer: Cham, Switzerland, 2024; Volume 1026. [CrossRef]
22. Moya-Salazar, J.; Rojas-Zumaran, V. Eco-Pap: The Ecological Modification of the Papanicolaou Stain for Sustainable Cervical Cancer Diagnosis. *Acta Cytol.* **2019**, *63*, 35–43. [CrossRef]
23. Mazumder, M.K.A.; Nobi, M.M.U.; Mridha, M.F.; Hasan, K.T. A Precise Cervical Cancer Classification in the Early Stage Using Transfer Learning-Based Ensemble Method: A Deep Learning Approach. In *Data-Driven Clinical Decision-Making Using Deep Learning in Imaging*; Springer: Singapore, 2024; Volume 152. [CrossRef]

24. Goswami, A.; Goswami, N.G.; Sampathila, N. Deep Learning-Based Classification of Cervical Cancer using pap smear images. In Proceedings of the 2024 4th International Conference on Intelligent Technologies CONIT 2024, Bangalore, India, 21–23 June 2024. [CrossRef]
25. Maurya, R.; Deshwal, S.; Rajput, L.; Gopalakrishnan, T.; Krishna, R.J. Cervical Cancer Classification using Multi-Scale, Multiple-Attention-based Deep Learning Approach. In Proceedings of the 2024 Asian Conference on Intelligent Technologies ACOIT 2024, Kolar, India, 6–7 September 2024. [CrossRef]
26. Babuc, D.; Fortiş, A.-E. A Customizable Intelligent System for Cervical Cytology Image Classifications. In *Complex, Intelligent and Software Intensive Systems*; Springer: Cham, Switzerland, 2024; Volume 87. [CrossRef]
27. Zhang, C.; Liu, D.; Wang, L.; Li, Y.; Chen, X.; Luo, R.; Che, S.; Liang, H.; Li, Y.; Liu, S.; et al. DCCL: A Benchmark for Cervical Cytology Analysis. In *Machine Learning in Medical Imaging*; Springer: Cham, Switzerland, 2019; Volume 11861. [CrossRef]
28. Sahoo, P.; Saha, S.; Mondal, S.; Sharma, S.K. Boosting cervical cancer detection with a multi-stage architecture and complementary information fusion. *Soft Comput.* **2025**, *29*, 1191–1206. [CrossRef]
29. Merlina, N.; Prasetyo, A.; Zuniarti, I.; Mayangky, N.A.; Sulistyowati, D.N.; Aziz, F. Improving Early Detection of Cervical Cancer Through Deep Learning-Based Pap Smear Image Classification. *J. Appl. Data Sci.* **2025**, *6*, 969–980. [CrossRef]
30. Ouh, Y.-T.; Kim, T.J.; Ju, W.; Kim, S.W.; Jeon, S.; Kim, S.-N.; Kim, K.G.; Lee, J.-K. Development and validation of artificial intelligence-based analysis software to support screening system of cervical intraepithelial neoplasia. *Sci. Rep.* **2024**, *14*, 1957. [CrossRef]
31. Chowdary, G.J.; Suganya, G.; Premalatha, M.; Yogarajah, P. Nucleus segmentation and classification using residual SE-UNet and feature concatenation approach incervical cytopathology cell images. *Technol. Cancer Res. Treat.* **2023**, *22*, 15330338221134833. [CrossRef]
32. Chauhan, N.K.; Kumar, A.; Singh, K.; Kolambakar, S.B. HDFCN: A Robust Hybrid Deep Network Based on Feature Concatenation for Cervical Cancer Diagnosis on WSI Pap Smear Slides. *BioMed Res. Int.* **2023**, *2023*, 4214817. [CrossRef]
33. Rezende, M.T.; Silva, R.; Bernardo, F.d.O.; Tobias, A.H.G.; Oliveira, P.H.C.; Machado, T.M.; Costa, C.S.; Medeiros, F.N.S.; Ushizima, D.M.; Carneiro, C.M.; et al. Cric searchable image database as a public platform for conventional pap smear cytology data. *Sci. Data* **2021**, *8*, 151. [CrossRef]
34. Kaur, H.; Kaur, J.; Sharma, R. Comparison of deep transfer learning models for classification of cervical cancer from pap smear images. *Sci. Rep.* **2025**, *15*, 3945. [CrossRef]
35. Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. DINOv2: Learning Robust Visual Features without Supervision. *Comput. Vis. Pattern Recognit.* **2024**, *2*. [CrossRef]
36. Alohali, M.A.; El-Rashidy, N.; Alaklabi, S.; Elmannai, H.; Alharbi, S.; Saleh, H. Swin-GA-RF: Genetic algorithm-based Swin Transformer and random forest for enhancing cervical cancer classification. *Front. Oncol.* **2024**, *14*, 1392301. [CrossRef]
37. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. *arXiv* **2021**. [CrossRef]
38. Rodriguez, J.; Huang, K.; Xu, M. Multi-Task Breast Ultrasound Image Classification and Segmentation Using Swin Transformer and VMamba Models. In *2024 7th International Conference on Pattern Recognition and Artificial Intelligence (PRAI)*; IEEE: Piscataway, NJ, USA, 2024; pp. 858–863. [CrossRef]
39. Rao, Y.; Zhao, W.; Zhu, Z.; Zhou, J.; Lu, J. GFNet: Global Filter Networks for Visual Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 10960–10973. [CrossRef]
40. Ahmed, H.O.A.; Nandi, A.K. Token Mixing for Breast Cancer Diagnosis: Pre-Trained MLP-Mixer Models on Mammograms. *IEEE Access* **2025**, *13*, 120190–120208. [CrossRef]
41. Prokhorenkova, L.; Gusev, G.; Vorobev, A.; Dorogush, A.V.; Gulin, A. CatBoost: Unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems 31*; Curran Associates, Inc.: New York, NY, USA, 2018.
42. Tan, S.L.; Selvachandran, G.; Ding, W.; Paramesran, R.; Kotecha, K. Cervical Cancer Classification from Pap Smear Images Using Deep Convolutional Neural Network Models. *Interdiscip. Sci. Comput. Life Sci.* **2024**, *16*, 16–38. [CrossRef]
43. Bhatt, A.R.; Ganatra, A.; Kotecha, K. Cervical cancer detection in pap smear whole slide images using convNet with transfer learning and progressive resizing. *PeerJ Comput. Sci.* **2021**, *7*, e348. [CrossRef]
44. Maurya, R.; Pandey, N.N.; Dutta, M.K. VisionCervix: Papanicolaou cervical smears classification using novel CNN-Vision ensemble approach. *Biomed. Signal Process. Control* **2023**, *79*, 104156. [CrossRef]
45. Khan, A.; Han, S.; Ilyas, N.; Lee, Y.-M.; Lee, B. CervixFormer: A Multi-scale swin transformer-Based cervical pap-Smear WSI classification framework. *Comput. Methods Programs Biomed.* **2023**, *240*, 107718. [CrossRef]
46. Kalbhor, M.; Shinde, S.; Wajire, P.; Jude, H. CerviCell-detector: An object detection approach for identifying the cancerous cells in pap smear images of cervical cancer. *Heliyon* **2023**, *9*, e22324. [CrossRef]
47. Khiruddin, K.; Mustafa, W.A.; Islam, M.A.; Jamaludin, K.R.; Alquran, H.; Rahman, K.S.A. Comparative analysis of cervical cancer classification of DPAGCHE-enhanced Pap smear images using convolutional neural network models. *PLoS ONE* **2025**, *20*, e0330103. [CrossRef]

48. Zhang, L.; Lu, L.; Nogues, I.; Summers, R.M.; Liu, S.; Yao, J. DeepPap: Deep Convolutional Networks for Cervical Cell Classification. *IEEE J. Biomed. Health Inform.* **2017**, *21*, 1633–1643. [CrossRef]
49. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2017; pp. 2261–2269. [CrossRef]
50. Liu, Y.; Tian, Y.; Zhao, Y.; Yu, H.; Xie, L.; Wang, Y.; Ye, Q.; Jiao, J.; Liu, Y. VMamba: Visual State Space Model. *arXiv* **2024**. [CrossRef]
51. Attallah, O. CerCan·Net: Cervical cancer classification model via multi-layer feature ensembles of lightweight CNNs and transfer learning. *Expert Syst. Appl.* **2023**, *229*, 120624. [CrossRef]
52. Niu, S.; Zhang, L.; Wang, L.; Zhang, X.; Liu, E. Hybrid feature fusion in cervical cancer cytology: A novel dual-module approach framework for lesion detection and classification using radiomics, deep learning, and reproducibility. *Front. Oncol.* **2025**, *15*, 1595980. [CrossRef]
53. Chieregato, M.; Frangiamore, F.; Morassi, M.; Baresi, C.; Nici, S.; Bassetti, C.; Bnà, C.; Galelli, M. A hybrid machine learning/deep learning COVID-19 severity predictive model from CT images and clinical data. *Sci. Rep.* **2022**, *12*, 4329. [CrossRef]
54. Rainio, O.; Teuho, J.; Klén, R. Evaluation metrics and statistical tests for machine learning. *Sci. Rep.* **2024**, *14*, 6086. Correction in *Sci. Rep.* **2024**, *14*, 15724. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.