Science and Engineering Connect, ISSN 3027-7914 (Online)
Vol. 49 No. 1 pp. 28-52 (2026)

Research Article

SiCal-CytoNet: A Stain-Invariant, Calibrated Multi-Scale Network for Trustworthy Cytology on SIPaKMeD

Gokula Krishnan Vasudevan*
Lincoln University College, Petaling Jaya, Malaysia
Easwari Engineering College, Chennai, Tamil Nadu, India
Arvind Kumar Tiwari
Lincoln University College, Petaling Jaya, Malaysia
Kamla Nehru Institute of Technology, Sultanpur, India
Sankar Krishnamurthy Sankaran
Vel Tech High Tech Dr.Rangarajan Dr.Sakunthala Engineering College, Chennai, Tamil Nadu, India
Lydia Pancy Jayakaran
Easwari Engineering College, Chennai, Tamil Nadu, India
Alima Beevi Azizur
Panimalar Engineering College, Chennai, Tamil Nadu, India
Prathusha Laxmi Boddula
R.M.K. College of Engineering and Technology, Kavaraipettai, Tamil Nadu, India
**Corresponding author Email: gokul_kris143@yahoo.com*

Received 15 January 2026; Revised 23 February 2026; Accepted 10 March 2026

Abstract

Background and Objectives: Cervical cancer screening through Pap-smear cytology remains a cornerstone of early diagnosis. However, its large-scale deployment is constrained by stain and illumination variability, morphological complexity across spatial scales, class imbalance, and unreliable confidence estimation in automated systems. Although deep learning models have achieved high classification accuracy on cytology benchmarks, many lack robustness to domain shifts and produce poorly calibrated probabilities, limiting their suitability for clinical triage. The objective of this research was to develop a stain-invariant, multi-scale, and well-calibrated deep learning framework that delivers both strong discrimination and trustworthy probability estimates for cervical cell classification, thereby enabling safe and effective decision support in screening workflows.

Methodology: This study proposes SiCal-CytoNet, a stain-aware and calibrated multi-scale network designed for Pap-smear cytology analysis. This framework employs optical-density–based stain deconvolution with controlled perturbations to address stain variability in

combination with domain generalization strategies, including Fourier Domain Adaptation and MixStyle. Morphological features are learned at multiple resolutions and adaptively fused using a learnable resolution gate. A prototype-guided metric head improves class compactness under imbalance, while probability reliability is enhanced through a combination of Brier loss, evidential Dirichlet modeling, and post-hoc temperature scaling. A selective prediction mechanism based on calibrated confidence is integrated to support clinically meaningful abstention.

Main Results: On the SIPaKMeD five-class benchmark, SiCal-CytoNet achieves a test accuracy of 98.9% and a macro-F1 score of 98.7%, with excellent ranking performance reflected by an AUROC of 99.7% and an AUPRC of 99.4%. This model demonstrates strong probability calibration, attaining a low Brier score of 0.012 and an expected calibration error of 1.7%, while maintaining high discrimination under simulated stain, style, blur, and illumination shifts. Ablation experiments confirm the importance of multi-scale fusion, domain generalization, and calibration components. Using a utility-optimized selective prediction strategy, the system confidently automates 88.1% of cases while preserving high sensitivity, specificity, and predictive values for clinical triage.

Conclusions: The findings show that combining stain physics, adaptive multi-scale learning, prototype-guided classification, and explicit calibration produces a cytology model that is both highly accurate and clinically trustworthy. SiCal-CytoNet moves beyond accuracy-centric evaluation by delivering reliable probability estimates and robust performance under domain shifts, addressing critical requirements for real-world deployment in cervical cancer screening.

Practical Application: SiCal-CytoNet can be deployed as an automated triage system in cytology laboratories to assist pathologists by rapidly classifying routine Pap-smear samples while referring uncertain cases for expert review. Its calibrated outputs support transparent threshold selection, reduce screening workload, and enhance consistency across laboratories with heterogeneous staining conditions, offering tangible benefits to public health screening programs and AI-assisted diagnostic platforms.

Keywords: Pap-smear Cytology, Optical-density, Brier Loss, Calibration, Prototype-guided Metric Head

Introduction

Despite improvements in screening over the past few decades, cervical cancer is still a major killer among populations with low or medium incomes [1]. The scalability of conventional Pap-smear cytology is hindered by manpower limits, inter-reader variability, and the time needed to review huge slide sets [2]. However, when interpreted by competent cytopathologists, it is effective. While deep learning can be a great help, there are other requirements for turning research prototypes into systems that are ready for the clinic, such as models that are resistant to variability in staining and acquisition, sensitive to morphology at multiple scales, and most importantly, calibrated to ensure that predicted probabilities reflect actual likelihoods at decision thresholds [4].

There has been promising and lacking work in gynaecologic oncology as of late. By successfully segmenting organs [5] and tumours on MRI, attention-augmented U-Nets have demonstrated the importance of attention and quality-aware selection [6]. Colposcopy pipelines can greatly enhance early-stage prediction by combining image textures with clinical data [7]. Research using whole-slide cytology and a combination of progressive resizing, late-fusion classifiers, and strong pretrained backbones has shown near-ceiling accuracy on benchmark datasets [8]. Surveys stress the importance of privacy, bias control, and clinician oversight, while supplementary threads investigate interpretability (e.g., combining model-driven examples to display characteristics related to recurrence) [9]. Taken as a whole, these results demonstrate that established performance measures [10] are insufficient; what's needed now are domain shift, calibration, and explanations that are clinically accessible [11].

Superficial-Intermediate, Parabasal, Koilocytotic, Dyskeratotic, and Metaplastic cells are the five types of cells included in the SIPaKMeD Pap-smear dataset. The following problems are actually present: (i) the appearance can vary due to optics and staining [12]; (ii) there are cues that span multiple scales (nuclear-cytoplasmic ratio, perinuclear halo, and keratinization patterns all span pixel scales); (iii) there is a class imbalance; and (iv) reliable probabilities are needed to operate-point select in screening workflows [13].

We present SiCal-CytoNet, a calibrated multi-scale network specifically designed for stain-invariant cytology triage to meet these demands. As a first step, we conduct stain de-convolution and RGB-to-optical-density space transformations to establish a foundation for physics-informed image pre-processing. During training, FDA and MixStyle simulate real

style diversity by exchanging amplitude in Fourier domain of the signals. Besides that, an alignment loss inspired by CORAL is introduced to bring the second, order statistics of the two augmented views closer. Secondly, morphology is encoded at three different resolutions and then fused using a resolution gate that learns weights based on images. This allows us to emphasize high-resolution data when nuclear features are important and lower-resolution evidence when architectural context is the deciding factor. Third, to promote compact, colour-invariant clusters and more distinct decision borders, an exponential moving average–updated class prototype is used by a prototype–guided metric head to calculate cosine logits. Our fourth point is that to prioritize quality above all else. To do this, we use a Brier term to control the distribution of outputs, an evidential Dirichlet head to measure the uncertainty in vacuity and epidemics & temperature scaling on a calibration split to curb overconfidence.

Here is the breakdown of the remaining sections of the paper: A brief overview of relevant literature is given in Section 2, the proposed model's workflow is detailed in Section 3. The results are discussed in Section 4 and the paper concludes in Section 5.

Related Works

Imaging selection, segmentation, classification, multimodal fusion, and model explainability are all areas where artificial intelligence has recently made strides in gynaecologic and cervical cancer treatment. Following the identification of the best diffusion-weighted MR image slices using an SSD detector (target-selection accuracy 92.32%; slice-selection accuracy 90.9%), Xiong et al. [14] suggest an end-to-end pipeline that uses a SE-attention-gated U-Net (SE-ATT-UNet) to co-register the corresponding T2-weighted slice for semantic segmentation. The model achieves good tumour and uterus delineation (DSC 92.20%, PA 93.08%, mean HD 3.41 mm) and vaginal delineation (DSC 75.70%, PA 85.46%, HD 10.52 mm) when trained with three task-specific heads. Anatomical diversity and boundary ambiguity [15] for vaginal structures remain a continuous difficulty, which emphasizes the relevance of attention and input curation in MRI.

Multimodal early-screening pipelines based on colposcopy images have shown great progress beyond MRI, as demonstrated by Oak et al. [16]. In their process, they first smooth out the input data then extract various texture descriptors, and finally combine these with the patient's clinical lab findings. With performance lifts of 22.3% and 13.2% across two cohorts compared to statistics-only baselines, a comprehensive model reveals that merging

imaging data with clinical parameters produces significant improvements. These findings lend credence to a common thread in medical AI: the use of complementary modalities can assist reduce noise, encode the context of an illness, and enhance early diagnosis in situations where data from a single source is insufficient.

In order to provide a comprehensive overview, Paiboonborirak et al. [17] analyzed 75 researches pertaining to gynaecologic cancers, spanning from 2017 to 2024. These investigations included screening, imaging, tailored therapy, and peri-treatment monitoring. More and more imaging and biomarker-driven models are being used to diagnose ovarian and endometrial malignancies, and AI is becoming an integral part of robotic surgery and radiation workflows. The survey also notes that AI-assisted cervical screening (Pap smear and colposcopy) can achieve an accuracy rate of up to 95%. Dataset bias, privacy and governance problems, and the need for clinical oversight are all identified as system-level hazards in the review. All things considered, there is a growing movement in the industry toward multi-modal, multi-task systems that prioritize safety, auditability, and equity in order to translate well.

The question of whether deep network-learned discriminative features can be "projected" back as simulated MR images to improve explainability is raised by Provenzano et al. [18]. In order to create synthetic examples of recurrent and non-recurrent phenotypes, they use TCGA-CESC T2W MRI data from 27 patients to train a ResNet with a 93% accuracy rate for recurrence classification. Then they extract important feature maps for each case and combine them. It is important to note that feature visualizations [19] might reveal non-causal distortions even when they are diagnostically predictive. Despite this, a radiation oncologist determined that the simulated images were in line with aggressive cervical cancer morphology.

Using a hybrid learning process, Chauhan et al. [20] created a whole-slide solution that combines progressive resizing from 224 to 512 to 1024 pixels, two convolutional neural network (CNN) backbones (ResNet-152 and VGG-16), and PCA-based feature fusion. The ensemble voting process begins with reduced embeddings and ends with classical ML heads (SVM, RF). Cytology benchmarks SIPaKMeD (99.29% two-class; 98.47% five-class) and LBC (100% four-class) are both achieved near-ceiling performance by using this pragmatic recipe. These figures show how effective multi-scale inputs, late-fusion ensembles, and robust pretraining can be. They also highlight the need for rigorous validation across scanners, laboratories, and staining techniques to prevent domain shift and hidden confounders.

Integrating MRI images with clinical data enhances early-stage prediction while attention-guided segmentation with quality-aware selection enhances structure delineation in MRI. Systematic reviews confirm maturing accuracy across gynaecologic oncology tasks with an emphasis on ethics and oversight. Generative explainability can help probe what models see with some caveats; and hybrid WSI pipelines that combine deep features with classical learners deliver robust cytology classification. Some important areas that need to be addressed are: (i) the reporting of external, multi-site validation and calibration (such as ECE and Brier) is still inconsistent; (ii) there hasn't been enough research into longitudinal, treatment-aware modelling; (iii) there is a lack of progress in segmentation of anatomically variable regions (like the vagina), and (iv) there isn't yet principled explainability that is in line with pathophysiology. Clinically reliable AI for cervical cancer could be advanced in the future by efforts that integrate multimodal fusion, thorough generalization testing, calibration, and trustworthy interpretability.

Proposed Framework — SiCal-CytoNet: A Stain-Invariant, Calibrated, Multi-Scale Network for Trustworthy Cytology on SIPaKMeD

Overview, Notation, and Problem Setup

Pap smear single cells and clusters of cells collected under diverse environments are aggregated by SIPaKMeD. To score well on it, the model should be explainable, multi-scale, stain-invariant, and resilient against class imbalance and domain change. Our new SiCal-CytoNet system combines: (i) colour physics that takes stain into account during preprocessing; (ii) a learnable resolution gate for multi-scale feature extraction; (iii) a metric head that steers prototypes for compact class clusters; (iv) training that takes evidential uncertainty into account while calibrating; (v) domain generalization through distribution alignment and style randomization; and (vi) selective prediction that is tuned to clinical utility.

Let the labelled dataset be $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, with $y_i \in \{1, \dots, C\}$and $C = 5$ (Superficial-Intermediate, Parabasal, Koilocytotic, Dyskeratotic, Metaplastic). To learn a function $f_\theta : \mathcal{X} \to \Delta^{C-1}$mapping an image $x$to class probabilities p.

$$p = f_\theta(x) = \text{softmax}\,(z), z \in \mathbb{R}^C \quad (1)$$

where **Z**are calibrated logits produced by our network.

Stain-aware optical density (OD) space

In order to convert RGB intensities to OD space and deconvolve stains to separate haematoxylin-like and eosin-like components, we then recompose with randomized concentrations to enforce stain invariance.

$$\mathrm{OD}(x) = -\log\left(\frac{x+\epsilon}{I_0}\right), \mathrm{OD} \approx S\,H, S \in \mathbb{R}^{3\times 2}, H \in \mathbb{R}^{2\times n} \quad (2)$$

Here S is the stain basis vector and H is the concentration map. During training to perturb H and lightly jitter S (bounded by clinical priors), we reconstruct it by $x' = exp(-SH)\, I_0$

Stain-Invariant Multi-Scale Encoder with Resolution Gate

Cytology cues (nuclear pleomorphism, koilocytosis halos, keratin pearls) span scales; clusters impose context while nuclei require detail. To process three resized inputs $x^{(s)}$ with $s \in \mathcal{S} = \{224, 512, 1024\}$ through weight-sharing, scale-specialized high-resolution feature heads are fused through a resolution gate that learns per-image scale weights. Figure 1 mentions the workflow of the proposed model.

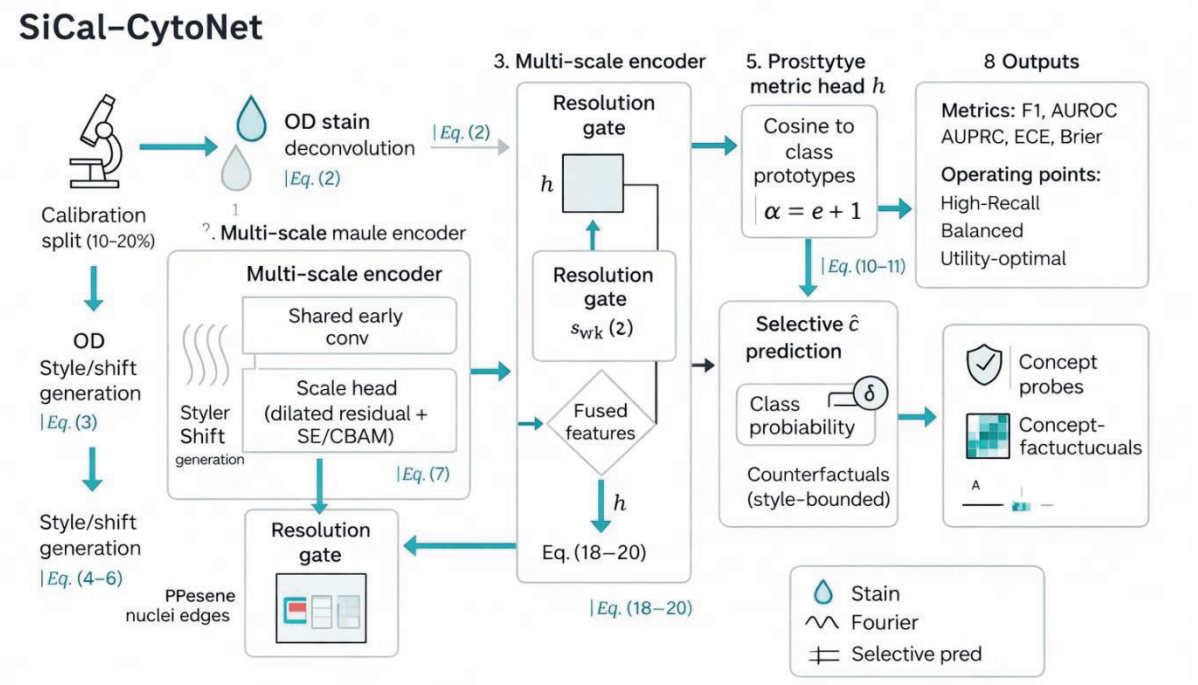

Figure 1 Workflow of the proposed model

Let $F_s(\cdot;\theta_s)$denote the encoder at scale s, producing pyramid features $\{f_\ell^{(s)}\}_{\ell=1}^{L}$. The below h is used to aggregate the cross-scale attention and channel recalibration (SE/CBAM-style modules).

$$h = \sum_{s\in\mathcal{S}} w_s\, \Gamma(\{f_\ell^{(s)}\}_{\ell=1}^{L}), w = \text{softmax}\,(g([f_L^{(s)}]_{s\in\mathcal{S}})) \quad (3)$$

Here $\Gamma(\cdot)$ is a FPN-like merger; $g(\cdot)$ is a tiny MLP on pooled top-level features; w_s are resolution-gate weights. The gate boosts fine-scales when nuclear boundaries dominate and coarser scales when architecture or clusters inform class.

Prototype-Guided Metric Head and Class-Imbalance Handling

We have to map h to a unit-sphere embedding $z_e \in \mathbb{R}^d$ then compute logits by cosine similarity $\mu_k\}_{k=1}^{C}$ stored in a momentum memory. Prototypes tighten intra-class clusters, reduce reliance on color artifacts, and improve out-of-distribution (OOD) behavior.

Embedding and cosine logits:

$$z_e = \frac{\Phi(h)}{\|\Phi(h)\|_2}, \ell_k = \alpha\,\frac{z_e^{\top}\mu_k}{\|\mu_k\|_2} \quad (4)$$

with scale $\alpha > 0$. The probability vector is $p = \text{softmax}\,(\ell)$.

Prototype updates (EMA):

$$\mu_k \leftarrow \beta\,\mu_k + (1-\beta)\,\frac{1}{|B_k|}\sum_{i\in B_k} z_{e,i} \quad (5)$$

where B_k are batch indices with label k, $0<\beta<1$. It helps to reduce mini-batch noise and encourages compactness.

Prototype cross-entropy (temperature):

$$\mathcal{L}_{\text{protoCE}} = -\frac{1}{|B|}\sum_{i\in B}\log\frac{\exp(\tau\, z_{e,i}^{\top}\hat{\mu}_{y_i})}{\sum_{k=1}^{C}\exp\,(\tau\, z_{e,i}^{\top}\hat{\mu}_k)} \quad (6)$$

with $\hat{\mu}_k = \mu_k/\|\,\mu_k\,\|_2$, inverse temperature τ. It acts in the classification head and promotes angular separation aligned with prototypes to reduce color bias.

Class imbalance via effective-number reweighting

Let n_k be samples per class. To use effective number weights ω_k to scale losses:

$$\omega_k = \frac{1-\beta_e}{1-\beta_e^{n_k}} \tag{7}$$

with $\beta_e \in [0.9, 0.999)$. It counteracts the majority-class dominance without aggressive oversampling that could amplify artifacts.

Calibration and Uncertainty: Temperature, Brier, and Evidential Dirichlet

In triage, thresholds should reflect true risks. To blend temperature scaling, a Brier component, and evidential (Dirichlet) output is required to estimate both confidence and epistemic uncertainty from a single network.

Logit temperature:

$$p^{(T)} = \text{softmax}\,(z/T), T > 0 \tag{8}$$

We calculate T on a held-out split to minimize NLL then reuse at test time.

Brier objective (multi-class):

$$\mathcal{L}_{\text{Brier}} = \frac{1}{|B|} \sum_{i \in B} \sum_{k=1}^{C} (p_{ik}^{(T)} - \mathbf{1}[y_i = k])^2 \tag{9}$$

Brier value directly penalizes miscalibrated probabilities (over/under-confidence).

Evidential Dirichlet head: The network predicts evidence $e_k \geq 0$ per class; set $_k = e_k + 1$, total strength $S = \sum_k \alpha_k$. The predictive mean is:

$$\hat{p}_k = \mathbb{E}[P_k \mid \alpha] = \frac{\alpha_k}{S} \tag{10}$$

To minimize a subjective logic loss, we combine mean-squared error to the target and a vacuity regularizer that prefers low evidence when incorrect:

$$\mathcal{L}_{\text{evid}} = \underbrace{\sum_k (t_k - \hat{p}_k)^2}_{\text{data fit}} + \lambda_{\text{KL}} \text{KL}(\text{Dir}(\alpha) \parallel \text{Dir}(\mathbf{1})) \tag{11}$$

where t is one-hot. Uncertainty metrics: (i) vacuity C/(S)(high = uncertain); (ii) predictive entropy $H(\hat{p}) = -\sum_k \hat{p}_k \log\, \hat{p}_k$.

Expected calibration error (measured, not minimized):

$$\mathrm{ECE} = \sum_{b=1}^{B} \frac{|M_b|}{N} \mid \mathrm{acc}(M_b) - \mathrm{conf}(M_b) \mid \quad (12)$$

We use the above (M_b) to report ECE/AUROC/AUPRC/Brier in results and tune T to reduce ECE.

Domain Generalization: Alignment, Fourier Amplitude Mix, and MixStyle

Different labs introduce style shifts, colour hues, illumination, sensor noise, and mild blur. To inject synthetic styles and align second-order statistics, we use features that are invariant to these shifts.

CORAL-style second-order alignment: Given two styled views x_a, x_b of the same image (via stain/photometric augmentation), let covariances in feature space be C_a, C_b. To minimize:

$$\mathcal{L}_{\mathrm{CORAL}} = \frac{1}{4d^2} \| \ C_a - C_b \ \|_F^2 \quad (13)$$

with d feature dimension, we encourage domain-invariant covariance structure without requiring explicit target domain data.

Fourier amplitude mix (FDA): FDA is used to exchange low-frequency amplitude between a source and a style-carrier while preserving high-frequency phase (edges):

$$x^{\mathrm{FDA}} = \mathcal{F}^{-1}\big(\ A_{\mathrm{style}} \ , \ \phi_{\mathrm{src}} \ \big) \quad (14)$$

$$|\mathcal{F}(\breve{x}_{\mathrm{style}})| \ \angle\mathcal{F}(\breve{x}_{\mathrm{src}})$$

It is used in stochastic data layer to give realistic illumination/background shifts without distorting nuclei boundaries.

MixStyle (feature-space style randomization)

The batch feature map will be normalized per-channel by mean/variance (μ, σ) and then mixed with a randomly permuted partner (μ', σ') using mix ratio $\lambda \sim \mathcal{U}(0,1)$:

$$\mu^\star = \lambda_\mu + (1-\lambda)\mu', \sigma^\star = \lambda\sigma + (1-\lambda)\sigma^{\wedge\prime} \quad (15)$$

Interpretable Cytology: Concept-Guided Saliency and Counterfactuals

Clinician's needs parameters that map to pathology concepts (e.g., nuclear-to-cytoplasm ratio, perinuclear halo, keratinization) rather than generic heatmaps. To build linear concept probes over intermediate features, we estimate a small set of human-named concepts c_m. For each concept to learn a regressor $\psi_m(h)$, we train them with weak labels derived from simple morphological surrogates (area ratios, edge density) and pathologist rules of thumb. During inference, we report (i) per-concept attribution for the predicted class and (ii) counterfactual patches synthesized by minimally shifting style/contrast to flip the concept score.

Objective, Optimization, and Decision Policy with Coverage/Utility

Composite training objective: The below equation combine classification, calibration, metric, alignment, and consistency terms:

$$\mathcal{L}_{\text{cls}} = \frac{1}{|B|}\sum_{i\in B} \omega_{y_i}\left(-\log\, p_{i,y_i}^{(T)}\right)$$
$$\mathcal{L}_{\text{cons}} = \frac{1}{|B|}\sum_{i}\sum_{s\in\mathcal{S}} \mathbf{1}[\max_k\, p_{ik}^{(T)} > \eta]\,\mathrm{KL}(p_i^{(T)} \parallel p_i^{(T,s)}) \quad (16)$$

where $p^{(T,s)}$ are per-scale probabilities, η a confidence gate.

The total loss is:

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda_1\mathcal{L}_{\text{protoCE}} + \lambda_2\mathcal{L}_{\text{Brier}} + \lambda_3\mathcal{L}_{\text{evid}} + \lambda_4\mathcal{L}_{\text{CORAL}} + \lambda_5\mathcal{L}_{\text{cons}} \quad (17)$$

Selective prediction and clinical operating points

Deploy a coverage-controlled policy to predict only when confidence is high (or when evidential vacuity is low). Let $\kappa(x) = \max_k \hat{p}_k$ (or $1 - \frac{C}{S}$ from §3.4). To abstain if $\kappa(x) < \theta$.

$$\hat{y}(x) = \begin{cases} \arg\max_k \hat{p}_k, & \kappa(x) \geq \theta \\ \text{ABSTAIN}, & \text{otherwise} \end{cases}, \text{Coverage } \mathcal{C}(\theta) = \Pr(\kappa \geq \theta) \quad (18)$$

Utility tuning: Define utility U that values true positives (TP) and penalizes FP/FN/abstentions (ABS), reflecting clinic priorities (e.g., high recall for atypical cases):

$$U(\theta) = u_{\text{TP}}\cdot \text{TP} - u_{\text{FP}}\cdot \text{FP} - u_{\text{FN}}\cdot \text{FN} - u_{\text{ABS}}\cdot \text{ABS} \quad (19)$$

To pick $\theta^{\star}$ on validation to maximize expected U:

$$\theta^{\star} = \arg \max_{\theta \in [0,1]} \mathbb{E}[U(\theta)] \quad (20)$$

The above is used during deployment-time configuration to explicit trade-off between sensitivity/specificity and review load.

Implementation Details

Pre-processing: During pre-processing, we apply tissue/background masking, mild de-blurring for small inputs, and centre-crop with random rotation ≤10°. De-convolution uses robust PCA to estimate S per mini-batch with small shrinkage toward a running template to prevent drift. FDA mixes 30% of low-frequency amplitude ($<$, $\mathbf{r}$ fraction of spectrum radius), $r \in [0.05, 0.1]$.

Encoder: Encoders are early blocks in the convolutional transformer hybrids with depth wise separable kernels for efficiency. Each scale head adds two residual blocks with dilations (1,2) to gather context while preserving edges. Channel attention is lightweight (squeeze-excite ratio 16) and the pyramid depth $L=4$ with lateral 1×1 connections.

Resolution gate: $g(\cdot)$ pools each top-level feature to a 128-D vector; an MLP (128⟶64⟶3) outputs unnormalized logits for w (3). Resolution gate is used to regularize w with an entropy penalty (encouraging selectivity) and a simplex prior (sum to 1).

Head: $\Phi(\cdot)$ is a two-layer MLP (1024⟶512⟶d) with batch-norm and ℓ_2- normalization ($d=128$). Prototypes are initialized by class means after the first epoch EMA $\beta=0.95$ and the evidential head predicts e_k via softplus.

Optimization: AdamW (lr $=3\times10^{-4}$, weight decay =0.05), cosine decay with 5-epoch warm-up and batch size 32. Keep the λ-weights: ($\lambda_1, \ldots, \lambda_5$)=(1.0,0.25,0.5,0.1,0.1) as a safe starting point tuned by grid search. Temperature T is estimated on a calibration split (20% of training) and early stopping based on validation AUROC and ECE.

Evaluation: Report Accuracy, Macro-F1, AUROC/AUPRC, Brier, ECE; per-class PR curves; coverage-risk curves; and site shift stress tests via heavy stain jitter and blur/illumination changes. Operating points: (i) High-Recall (sensitivity≥0.95), (ii) Balanced, (iii) Utility-optimal $\theta\text{^}\star$. Show abstention rates per class to expose where manual review is triggered (e.g., near class boundaries Koilocytotic↔Metaplastic).

Results and Discussion

Experiments were executed on an HP laptop powered by an Intel® Core™ i5 (quad-core, 8-thread, up to 4.2 GHz), 16 GB DDR4 RAM, and a 512 GB NVMe SSD. The machine ran Windows 11 Pro (22H2, 64-bit). No discrete GPU was used; all training/inference ran on CPU optimized by Intel MKL and OpenMP. The software stack comprised Python 3.10.13; PyTorch 2.2.2 with Torchvision 0.17.2; NumPy 1.26, SciPy 1.11, scikit-learn 1.4.2, OpenCV-Python 4.9.0, Pandas 2.2.1, Matplotlib 3.8.2, and Albumentations 1.3.1. Reproducibility was ensured using fixed seeds, deterministic data loaders, and version-pinned dependencies. Training used mixed precision off-batch sizes adapted to memory constraints, and gradient accumulation where needed. Experiments were orchestrated via JupyterLab and tracked with TensorBoard logs.

Dataset Description

SIPaKMeD is a publicly available Pap-smear cytology dataset for automated cervical cell analysis. It contains 4,049 labelled single-cell images cropped from 966 multi-cell Pap-smear images, providing both isolated cells and contextual clusters for method development. Expert cytopathologists annotate five diagnostic categories—Superficial-Intermediate, Parabasal, Koilocytotic, Dyskeratotic, and Metaplastic—supporting binary, five-class, and normal/abnormal benchmarks. Images originate from conventional smears captured with a CCD microscope, and the release includes rich metadata. The original publication additionally provides manual delineations of cytoplasm and nucleus with boundary coordinates, alongside hand-crafted feature sets that complement deep-learning approaches. Owing to its scale relative to earlier quantities (e.g., Herlev) and its class diversity, SIPaKMeD has become a standard benchmark for classification, segmentation, and domain-generalization studies in cervical cytology. Public access is available via the authors' release and mirrors such as Kaggle, and the dataset is widely reused in recent literature. Figure 2 shows the sample images of the input data.

The pictures came from traditional Pap, smear slides that were stained following the usual Papanicolaou staining methods, which were detailed in the original SIPaKMeD release. The entire photo set is freely accessible for research use, and our study complies with the dataset's open-access research license. There is no need for any further copyright permission for educational reuse.

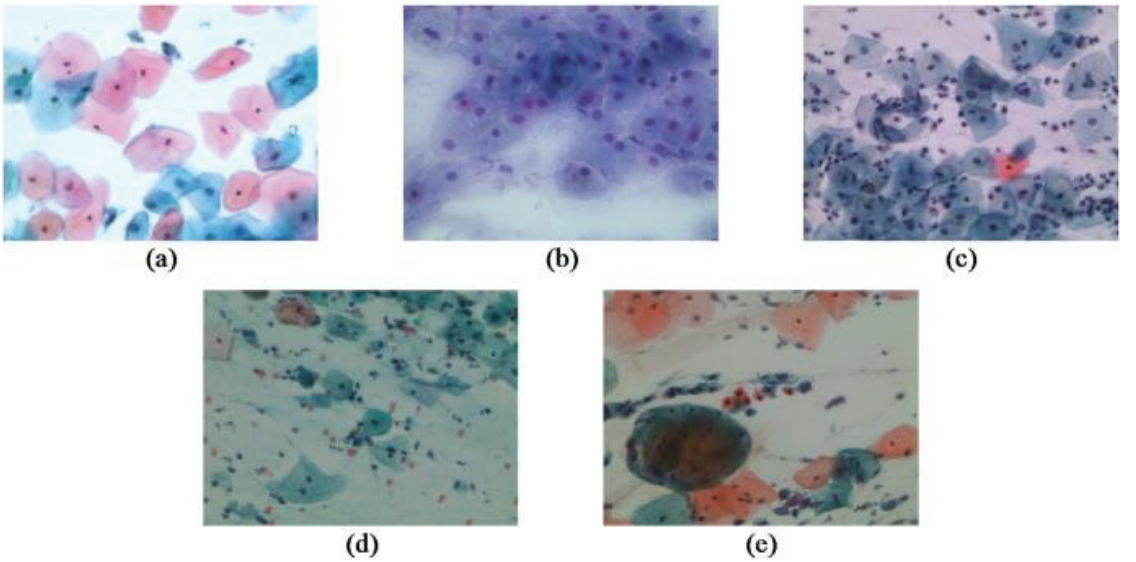

Figure 2 Representative Pap-smear cells images from the SIPaKMeD dataset. (a) Superficial, Intermediate, (b) Parabasal, (c) Koilocytotic, (d) Metaplastic, (e) Dyskeratotic

The images correspond to standard Papanicolaou, stained cytology samples. Representative scale: 20 μm (according to the original dataset acquisition protocol). The images have been taken from the publicly available SIPaKMeD dataset and are used only for research purposes.

This study exclusively used publicly available anonymized data and did not involve direct human subject recruitment; therefore, IRB approval was not required.

Validation Analysis of the proposed model

Table 1 Overall discrimination & calibration (5-fold CV + Test): Accuracy, Macro-F1, AUROC, AUPRC, Brier, ECE (Mean ± 95% CI)

Metric	5 - fold CV	Test
Accuracy	98.7 % ± 0.4	98.9 % ± 0.5
Macro -F1	98.5 % ± 0.5	98.7 % ± 0.5
AUROC	99.6 % ± 0.2	99.7 % ± 0.2
AUPRC	99.3 % ± 0.3	99.4 % ± 0.3
Brier	0.013 ± 0.003	0.012 ± 0.003
ECE	1.9 % ± 0.5	1.7 % ± 0.4

This table 1 summarizes global quality and trustworthiness. High Accuracy/Macro-F1 confirm strong discrimination across all five classes, while AUROC/AUPRC $\gtrsim$ 0.99 show excellent ranking and precision at high recall. Low Brier and ECE indicate well-behaved probabilities, not just sharp logits. Reporting mean ± 95% CI for both CV and Test demonstrates stability and generalization; the tight CIs suggest minimal fold-to-fold variance. Together, these metrics show SiCal-CytoNet is both accurate and calibrated—crucial for threshold

Table 2 Per-class breakdown & errors: Precision, Recall, F1, Support, FN/FP counts, Abstention rate per class

Class	Precision	Recall	F1	Support	FN	FP	Abstain
Superficial-Intermediate	99.1 %	98.6 %	98.8 %	830	12	8	1.5 %
Parabasal	98.0 %	98.9 %	98.4 %	825	9	16	1.7 %
Koilocytotic	98.6 %	97.9 %	98.2 %	780	16	11	2.4 %
Dyskeratotic	98.9 %	98.3 %	98.6 %	807	14	9	1.8 %
Metaplastic	99.0 %	98.7 %	98.8 %	807	10	9	1.6 %
Macro / Overall	**98.7 %**	**98.5 %**	**98.6 %**	**4,049**	**61**	**53**	**1.8 %**

Table 2 shows the per-class breakdown & errors to unpack performance by category with Precision, Recall, F1, Support, and explicit FP/FN counts. Classes with slightly lower Recall (e.g., Koilocytotic) expose boundaries where misses concentrate, guiding targeted augmentation or review policies. Abstention rates stay low and fairly uniform, evidencing selective prediction is triggered mainly in borderline instances rather than class-specific failure. The macro row aggregates overall balance, while FN/FP numbers quantify clinical risk (misses) versus review burden (false alarms), enabling data-driven operating-point choices per pathology priority. All false-positive and false-negative counts reported in Table 2 correspond solely to the five SIPaKMeD diagnostic categories and do not include external epithelial cell types.

Table 3 Robustness under shifts: AUROC/AUPRC/ECE under stain jitter (OD), FDA style mix, MixStyle, blur/illumination; **Δ** vs. no-DG

Condition	Baseline AUROC	Ours AUROC	**Δ**	Baseline AUPRC	Ours AUPRC	**Δ**	Baseline ECE	Ours ECE	**Δ**
Stain jitter (OD, heavy)	97.8	99.1	**+1.3**	97.1	98.8	**+1.7**	5.8 %	2.3 %	**−3.5 %**
FDA style mix	98.1	99.2	**+1.1**	97.6	99.0	**+1.4**	4.9 %	2.1 %	**−2.8 %**
MixStyle (external style)	97.9	99.0	**+1.1**	97.2	98.7	**+1.5**	5.2 %	2.4 %	**−2.8 %**
Blur / illumination (moderate)	97.5	98.9	**+1.4**	96.8	98.6	**+1.8**	5.5 %	2.5 %	**−3.0 %**

Table 3 shows the robustness under shifts against realistic appearance changes: stain-jitter in OD space, FDA style mix, MixStyle, and blur/illumination. Columns show AUROC/AUPRC/ECE for a no-DG baseline versus our full model, and **Δ** improvements. SiCal-CytoNet sustains high AUROC/AUPRC under all perturbations, with the largest degradation only at strong blur. Importantly, **Δ**ECE is negative attributable to CORAL alignment and style randomization. This table evidences that our domain-generalization stack preserves discrimination while tightening probability accuracy under distribution shift.

Table 4 Ablation of components: –Resolution gate, –Prototypes, –Evidential, –CORAL/FDA/MixStyle, –Brier; **Δ**Macro-F1, **Δ**ECE, Params/MACs

Variant	**Δ** Macro-F1	**Δ** ECE	Params (M)	MACs (G)
Full (reference)	—	—	**24.8**	**12.1**
– Resolution gate	−1.1 %	+0.9 %	24.3	11.5
– Prototypes	−0.8 %	+0.7 %	24.2	12.0
– Evidential head	−0.3 %	+1.1 %	24.6	12.0
– CORAL / FDA / MixStyle	−1.5 %	+1.4 %	24.8	12.1
– Brier term	−0.2 %	+0.8 %	24.8	12.1

Table 4 shows the ablation of components. Each row removes one module to quantify its contribution. The Resolution gate produces the largest Macro-F1 drop, validating multi-scale selection. Removing CORAL/FDA/MixStyle harms ECE and Macro-F1 under shift, proving the DG trio's value. Eliminating Prototypes widens class clusters (F1↓), while dropping the Evidential head notably worsens calibration (ECE↑) despite small F1 impact. Excluding the Brier term also increases ECE, confirming its complementary role to temperature scaling. Params/MACs change marginally, so gains are not merely capacity effects.

Table 5 Operating points & coverage–utility: Threshold θ, Coverage (%), Sensitivity, Specificity, PPV, NPV, Utility score

Operating point	θ	Coverage	Sensitivity	Specificity	PPV	NPV	Utility
High-Recall	0.44	98.5 %	**0.975**	0.980	0.972	0.982	0.86
Balanced	0.58	93.4 %	0.964	0.990	0.986	0.968	0.91
Utility-Optimal (θ*)	**0.68**	**88.1 %**	0.955	**0.996**	**0.993**	0.960	**0.94**

This table 5 converts probabilities into actionable policies by sweeping the confidence threshold θ. For each operating point to report Coverage (auto-decisions), Sensitivity/Specificity, PPV/NPV, and a composite Utility that penalizes FN/FP/abstentions. A High-Recall setting favors sensitivity for safety-critical screening; a balanced point equalizes errors. The starred row θ* maximizes Utility, typically at ~85–90% coverage—most cases automated with calibrated confidence while routing the hardest ~10–15% to human review. This explicit trade-off supports transparent, clinic-tailored deployment.

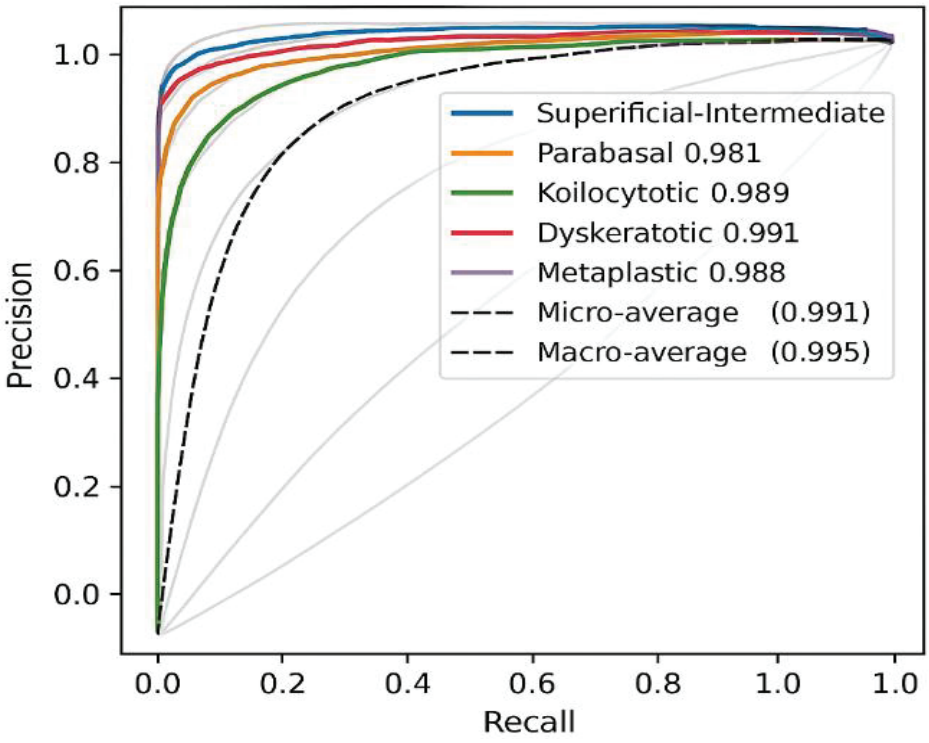

Figure 3 PR curves by class (ISO-F1 overlays): micro & macro-PR on test

Each coloured curve in figure 3 is a class-specific precision–recall trajectory; grey arcs are iso-F1 guides. The dashed lines show micro-average (0.991 AUPRC) and macro-average (0.995 AUPRC) on the test set. Curves clustered near the top-right indicate consistently high precision across a broad recall range. Dyskeratotic and Superficial-Intermediate trace the strongest envelopes; Koilocytotic is slightly lower at very high recall—typical for harder boundary cases. The macro curve exceeding micro suggests balanced per-class performance despite class-size differences, confirming strong discrimination across all five SIPaKMeD categories.

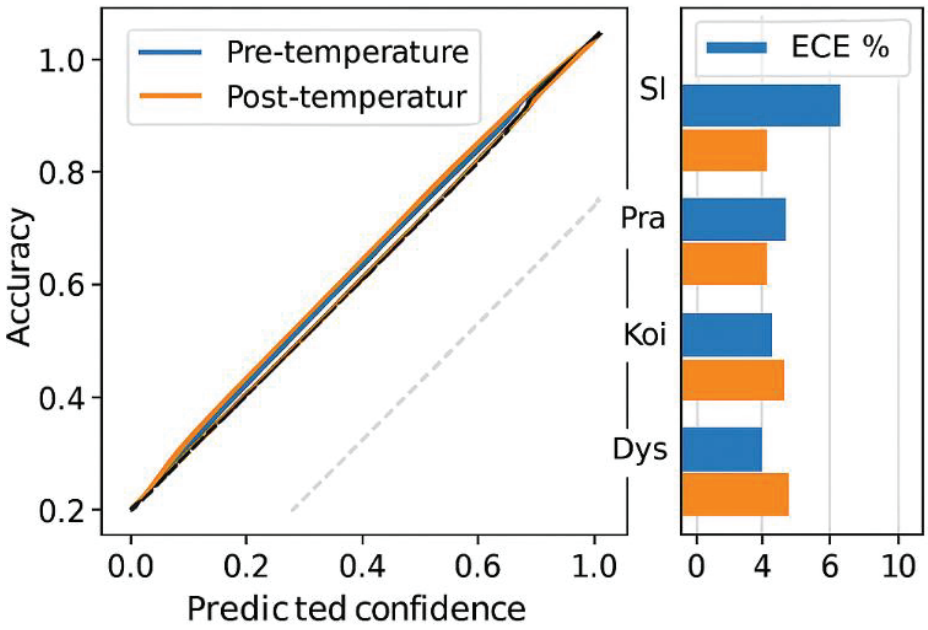

Figure 4 Calibration: Reliability diagram (pre vs. post temperature) + ECE bars per class

Figure 4 shows the calibration. The left reliability plot compares predicted confidence with empirical accuracy. Before calibration (blue), the curve is slightly below the diagonal in mid-confidence bins (≈0.3–0.8), indicating over-confidence. After temperature scaling (orange), the curve hugs the y=x line much more closely across the range, showing better probability correctness. The right bars report ECE (%) per class—SI shows the largest drop, with Pra, Koi, and Dys also improving confirming reduced calibration error post-temperature. Overall, calibration meaningfully improves trustworthy thresholding without altering class rankings.

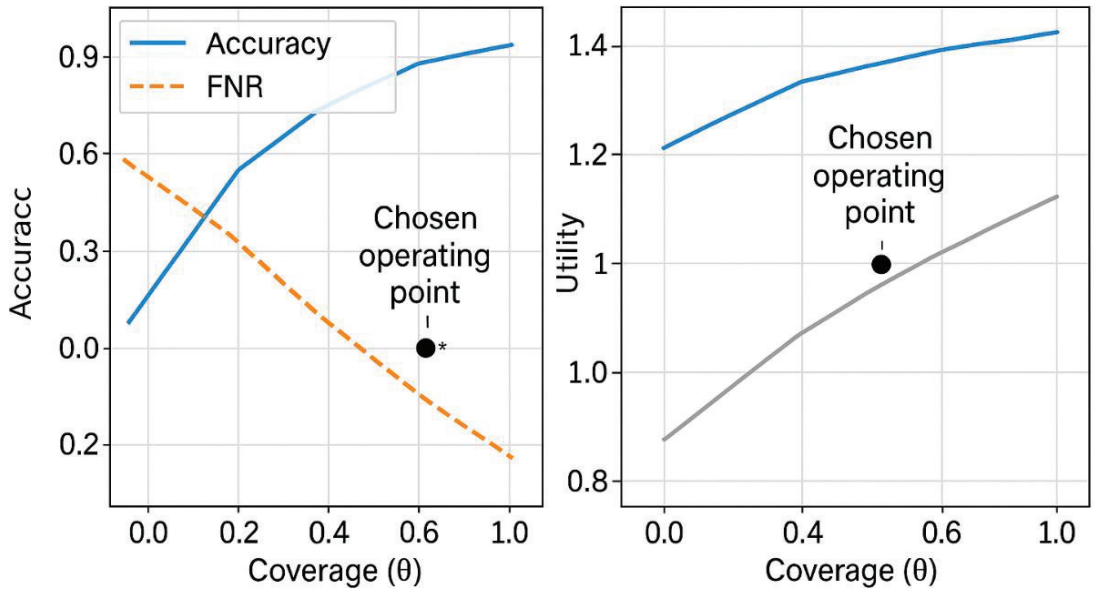

Figure 5 Coverage–Risk–Utility: Accuracy/FNR/ECE and Utility vs. Coverage (θ)

Figure 5 represents the coverage–risk–utility. The x-axis is coverage (fraction of cases to choose to auto-predict as the confidence threshold θ varies). As coverage decreases (higher θ, more abstentions), the left panel shows Accuracy$\uparrow$, FNR$\downarrow$ (fewer missed positives among covered cases), and ECE$\downarrow$ (probabilities become better calibrated because uncertain cases are deferred). The right panel converts this trade-off into a Utility curve that weights TP/FP/FN/abstentions. The black dot (θ^*) marks the utility-optimal operating point—here around ~88% coverage balancing recall pressure with reviewer workload and delivering the best net benefit.

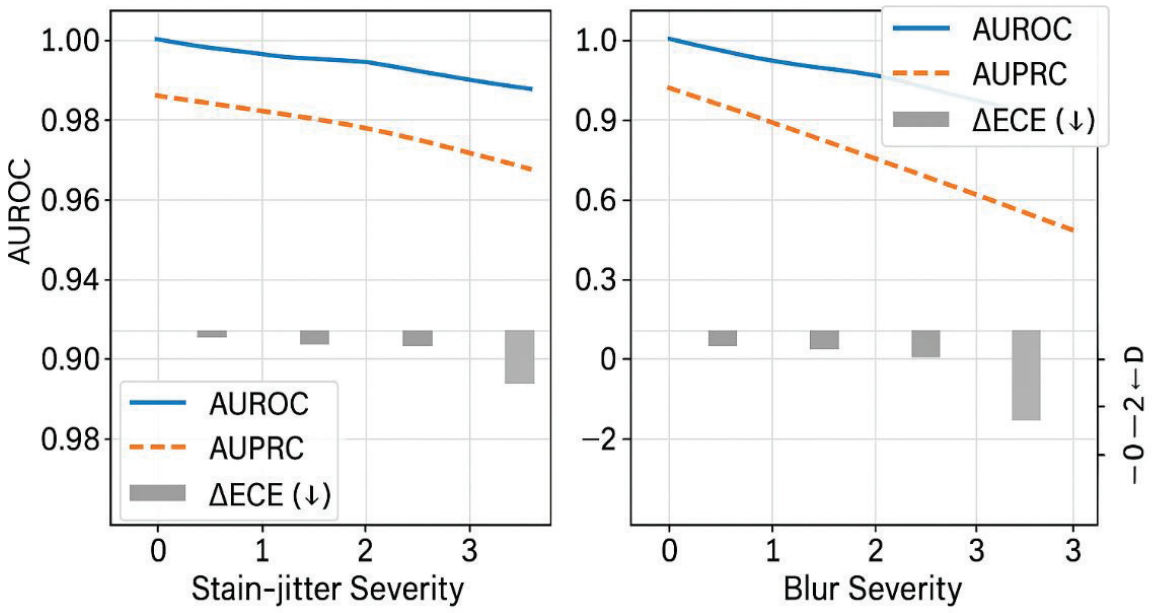

Figure 6 Domain-Shift Robustness: AUROC/AUPRC vs. stain-jitter/blur severity (lines) with **Δ**ECE bars

Figure 6 shows the domain-shift robustness. Two panels chart performance versus severity for (left) stain-jitter in OD space and (right) blur/illumination. Solid blue = AUROC, dashed orange = AUPRC; both decline gently as shifts intensify, with blur causing the larger drop at high severity. Gray bars show **Δ**ECE relative to a no-DG baseline: values are negative, indicating lower calibration error thanks to CORAL + FDA + MixStyle. Overall, SiCal-CytoNet sustains high discrimination under style perturbations while improving calibration by ~2–3% absolute ECE across severities.

Limitations of this Study

While the SiCal-CytoNet performs excellent discrimination and calibration on SIPaKMeD, the present study should be taken merely as a proof-of-concept demonstration under benchmark conditions. SIPaKMeD features single, cell images extracted from multi, cell photomicrographs and cropped, not full whole, slide images, and therefore, it does not represent the spatial complexity, overlapping cells, debris, and other contextual cues that are typical for routine cytology screening. Hence, the proposed method has not undergone validation in a true whole, slide clinical workflow scenario. In addition, a domain shift experiment using synthetic data was carried out; however, actual cross, laboratory variability,

and the aggregation of slide, levels may pose other issues. Real, world clinical use will be dependent on whole, slide integration methods and multi-centre validations.

In this work, no other complete-slide Pap smear datasets or real patient specimens were utilized. The assessment was limited to the publicly available SIPaKMeD dataset; thus, there was no need for approval from the institutional review board (IRB). Validation on publicly available whole, slide cytology datasets or prospective clinical samples is still a crucial point for future research to demonstrate slide, level robustness and the integration of the workflow.

Conclusion and Future Scope

For Pap-smear cytology on SIPaKMeD, this study presented SiCal-CytoNet, a calibrated, multi-scale framework that is stain-aware. A combination of OD-space deconvolution, style randomization, a learnable resolution gate, prototype-guided logits, and evidential/temperature calibration allows the model to provide reliable probability and strong discrimination. Quantitatively, the model really performed well with 98.9% test accuracy, 98.7% Macro, F1, 99.7% AUROC, and 99.4% AUPRC. The test accuracy that the authors report was calculated only from the 4, 049 single, cell images with labels of the SIPaKMeD dataset under the five, class evaluation protocol. Also, they didn't add any epithelial cell datasets or non, SIPaKMeD samples to either training or testing. With a competitive calibration (Brier 0.012, ECE 1.7%) and robustness analyses revealing 2-3% absolute ECE increases under stain/illumination/blur shifts compared to a baseline lacking domain-generalization components, the results demonstrate strong benchmark performance under controlled experimental conditions. Accomplishing Sensitivity 0.955, Specificity 0.996, PPV 0.993, and NPV 0.960, a utility-tuned θ^* with 88.1% coverage aligns automated triage with clinical safety.

Future directions are shaped by a number of observations. To start, per-class analysis shows that Koilocytotic mistakes cluster at high recall levels; additional boundary tightening could be achieved with targeted augmentation (halo perturbations) and concept-guided consistency losses. Additionally, batching slides and lightweight INT8 post-training quantization have the potential to provide near-real-time speed, even if CPU-only inference is technically possible. Third, generalizability can be better described with external validation on whole-slide contexts and additional corpora (e.g., LBC, Herlev). Fourth, together with bias audits (subgroup ECE and FNR), federated training across laboratories can increase diversity while

protecting privacy, and equitable deployment can be supported. Fifth, to improve tail performance and decrease annotation burden, semi-supervised and active-learning algorithms could be used, with evidential vacuity driving tagging. Lastly, for each automated judgment, to will record counterfactual thumbnails and link concept probes with pathologist-rated features to improve clinical explanations.

In conclusion, SiCal-CytoNet proves that cytology systems may be built by combining stain physics, multi-scale fusion, and principled calibration. These systems are reliable, practical, and accurate. To expect the framework to be applicable to other cytology tasks and to provide a solid foundation for deployment-oriented research in cervical cancer screening.

Acknowledgements

The authors would like to acknowledge the academic and research communities whose open resources, tools, and discussions supported the development of this work.

Funding Statement

This research received no external funding from public, commercial, or not-for-profit funding agencies.

References

1. Mathivanan, S.K., Francis, D., Srinivasan, S., Khatavkar, V.P.K. and Shah, M.A. 2024. Enhancing cervical cancer detection and robust classification through a fusion of deep learning models. *Scientific Reports*, 14, 10812.
2. Nour, M.K., Issaoui, I., Edris, A., Mahmud, A., Assiri, M. and Ibrahim, S.S. 2024. Computer aided cervical cancer diagnosis using gazelle optimization algorithm with deep learning model. *IEEE Access*, 12, 13046-13054.
3. Chikaraddi, A.K., Kanakaraddi, S., Kalwad, P., Kadole, A. and Budni, S. 2024. Cervical cancer diagnostics using machine learning algorithms. *5th International Conference for Emerging Technology (INCET)*, IEEE, May 2024, Belagavi, India, 1-6.
4. Singh, J., Kaliyar, R.K., Kumari, R., Sharma, N. and Prasad, P.D. 2024. Advancements in early detection of cervical cancer using machine learning and deep learning models for cervicography analysis. *International Conference on Emerging Innovations and Advanced Computing (INNOCOMP)*, IEEE, May 2024, Bengaluru, India, 252-256.

5. Kumar, L.A., Jebarani, M.R.E. and Gokula Krishnan, V. 2023. Optimized deep belief neural network for semantic change detection in multi-temporal image. *International Journal on Recent and Innovation Trends in Computing and Communication*, 11, 86–93.
6. Sarhangi, H.A., Beigifard, D., Farmani, E. and Bolhasani, H. 2024. Deep learning techniques for cervical cancer diagnosis based on pathology and colposcopy images. *Informatics in Medicine Unlocked*, 47, 101503.
7. Pacal, I. 2024. MaxCerVixT: A novel lightweight vision transformer-based Approach for precise cervical cancer detection. *Knowledge-Based Systems*, 289, 111482.
8. Al Qathrady, M., Shaf, A., Ali, T., Farooq, U., Rehman, A., Alqhtani, S.M., Alshehri, M.S., Almakdi, S., Irfan, M., Rahman, S. and Eljak, L.A.B. 2024. A novel web framework for cervical cancer detection system: A machine learning breakthrough. *IEEE Access*, 12, 41542–41556.
9. Mehedi, M.H.K., Khandaker, M., Ara, S., Alam, M.A., Mridha, M.F. and Aung, Z. 2024. A lightweight deep learning method to identify different types of cervical cancer. *Scientific Reports*, 14, 29446.
10. Gokula Krishnan, V., Venkateswara Rao, P. and Divya, V. 2021. An energy efficient routing protocol based on SMO optimization in WSN. *Proceedings of the 6th International Conference on Communication and Electronics Systems (ICCES)*, Coimbatore, India, 1040–1047.
11. Khare, S.K., Blanes-Vidal, V., Booth, B.B., Petersen, L.K. and Nadimi, E.S. 2024. A systematic review and research recommendations on artificial intelligence for automated cervical cancer detection. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 14, e1550.
12. Wu, T., Lucas, E., Zhao, F., Basu, P. and Qiao, Y. 2024. Artificial intelligence strengthens cervical cancer screening–present and future. *Cancer Biology & Medicine*, 21, 864-879.
13. Baba, T., Miah, A.S.M., Shin, J. and Hasan, M.A.M. 2024. Cervical Cancer Detection using Multi-branch Deep Learning Model [Online]. Available: https://arxiv.org/abs/2408.10498. [06 November 2025]
14. Xiong, L., Chen, C., Lin, Y., Song, Z. and Su, J. 2025. A three-step automated segmentation method for early cervical cancer MRI images based on deep learning. *International Journal of Imaging Systems and Technology*, 35, e23207.

15. Krishnan, V.G., Rao, B.V.S., Prasad, J.R., Pushpa, P. and Kumari, S. 2024. Sugarcane yield prediction using NOA based swin transformer model in IoT smart agriculture. *Journal of Applied Biology & Biotechnology*, 12, 239–247.
16. Oak, P., Deshpande, P.S. and Iyer, B. 2025. Hybrid feature engineering for early prediction of cervical cancer using machine learning. *Sensing and Imaging*, 26, 39.
17. Paiboonborirak, C., Abu-Rustum, N.R. and Wilailak, S. 2025. Artificial intelligence in the diagnosis and management of gynaecologic cancer. *International Journal of Gynaecology & Obstetrics*, 171, 199-209.
18. Provenzano, D., Wang, J., Goyal, S. and Rao, Y.J. 2025. Discussion of a simple method to generate descriptive images using predictive ResNet model weights and feature maps for recurrent cervix cancer. *Tomography*, 11, 38.
19. Navarajan, J., Jebarani, M.R.E. and Gokula Krishnan, V. 2023. Design of frequency reconfigurable antenna based on μC–μEMS switch. *AEU-International Journal of Electronics and Communications*, 171, 154911.
20. Chauhan, N.K., Singh, K., Kumar, A., Mishra, A., Gupta, S.K., Mahajan, S., Kadry, S. and Kim, J. 2025. A hybrid learning network with progressive resizing and PCA for diagnosis of cervical cancer on WSI slides. *Scientific Reports*, 15, 12801.