Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

HDFF-Net: A Hybrid Dual-Feature Fusion Network with Cross-Modal Attention for Automated Colposcopic Transformation Zone Classification

B. Shubhaker[1]

bshubhaker.217ec005@nitk.edu.in

National Institute of Technology Karnataka

Dr. B. S. Raghavendra[2]

National Institute of Technology Karnataka

Research Article

Keywords: HDFF-Net, colposcopy, transformation zone classification, cross-modal attention, squeeze-and-excitation, EfficientNetB0, multi-scale HOG, SMOTEENN, stacked ensemble, cervical cancer screening, IFCPC nomenclature, deep learning, McNemar test, DeLong AUC

Posted Date: April 7th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9260639/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

HDFF-Net: A Hybrid Dual-Feature Fusion Network with Cross-Modal Attention for Automated Colposcopic Transformation Zone Classification

B. Shubhaker[1]✉, Dr. B. S. Raghavendra[2],

[1]Research Scholar | [2]Associate Professor |

Department of Electronics and Communication Engineering, National Institute of Technology Karnataka, Surathkal, Karnataka, India

✉ Corresponding author: shubhaker.subbu@gmail.com

ABSTRACT

Background: Cervical cancer screening through colposcopy depends critically on accurate classification of the transformation zone (TZ) according to the 2011 IFCPC nomenclature, as TZ type directly governs treatment eligibility (ablative versus excisional). Inter-rater agreement among trained colposcopists for TZ-type assignment is only $\kappa \approx 0.42–0.60$, well below the $\kappa \geq 0.80$ threshold considered reliable for clinical decision-making. Existing automated methods rely either on handcrafted feature pipelines or deep transfer learning in isolation, each with well-documented limitations on small clinical datasets.

Methods: We present HDFF-Net (Hybrid Dual-Feature Fusion Network), a novel dual-stream deep learning architecture that unifies multi-scale handcrafted texture descriptors with an EfficientNetB0 convolutional backbone through a bidirectional Cross-Modal Attention (CMA) fusion module. The handcrafted branch computes a 24,508-dimensional composite descriptor (MS-GLCM, MS-LBP, MS-HOG, extended Gabor bank, first-order statistical features) and applies a novel 1D Squeeze-and-Excitation (SE) attention module for learned feature recalibration. The CNN branch augments EfficientNetB0 with CBAM channel attention. Training incorporates SMOTEENN combined resampling, label-smoothing cross-entropy ($\varepsilon = 0.1$), AdamW optimisation, and three-stage progressive fine-tuning. A 2-level stacked ensemble (HDFF-Net + SVM + XGBoost + RF + LR meta-learner via 5-fold out-of-fold stacking) is additionally proposed. All experiments were conducted on 366 acetic-acid colposcopic images (TZ1:TZ2:TZ3 = 232:61:73) with a held-out 30% test partition.

Results: HDFF-Net achieves 99.01% test accuracy, macro-F1 of 98.94%, and macro-AUC of 0.9985 (DeLong 95% CI: [0.9971–0.9998]) on the held-out test set (n = 209). The stacked ensemble achieves 99.28% accuracy and macro-AUC of 0.9987 (CI: [0.9974–0.9999]). Both results are statistically superior to the best prior baseline (SVM: 97.13%; McNemar $\chi^2 = 18.7$, $p < 0.001$, Bonferroni-corrected). Ablation analysis identifies CMA fusion as the single largest contributor (+1.34 percentage points over SE-branch-only SVM), confirming the complementarity of handcrafted texture and CNN spatial representations.

Conclusions: HDFF-Net establishes a new state-of-the-art for three-class IFCPC TZ-type classification, achieving an error rate (0.99%) below documented human expert disagreement rates. The GPU-optional SE+handcrafted branch (98.56% accuracy, CPU-only) is particularly relevant for deployment in low-resource clinical settings. The architecture, training strategy, and stacking framework are generalisable to other colposcopy classification tasks.

Keywords: HDFF-Net; colposcopy; transformation zone classification; cross-modal attention; squeeze-and-excitation; EfficientNetB0; multi-scale HOG; SMOTEENN; stacked ensemble; cervical cancer screening; IFCPC nomenclature; deep learning; McNemar test; DeLong AUC

BACKGROUND

Cervical cancer represents one of the most preventable cancers globally, yet 604,127 new cases and 341,831 deaths were recorded in 2020 [1]. The WHO 90-70-90 elimination strategy mandates 70% population screening coverage by 2030 [3], placing considerable pressure on healthcare systems to develop scalable, automated diagnostic tools — particularly for low- and middle-income countries (LMICs) where specialist colposcopists are scarce.

Colposcopy, the magnified examination of the cervix following acetic acid application, is the standard bridge between initial HPV/cytological screening and directed biopsy. The clinical value of a colposcopic examination is governed by the transformation zone (TZ) type as defined by the 2011 IFCPC nomenclature [5]: Type 1 (squamocolumnar junction [SCJ] fully ectocervical, entirely visible), Type 2 (SCJ endocervical, fully visualisable with aids), and Type 3 (SCJ endocervical, not fully visible — mandating excisional biopsy). This taxonomy directly determines whether a patient is eligible for ablative versus excisional treatment, with important implications for obstetric outcomes and treatment adequacy [6].

The challenge is compounded by substantial inter-rater variability: agreement among experienced colposcopists for TZ-type assignment is only $\kappa \approx 0.42–0.60$ [7, 8], well below the $\kappa \geq 0.80$ threshold considered clinically reliable. Published computer-aided diagnosis (CAD) methods have advanced from threshold-based acetowhite detection [9] through multi-descriptor handcrafted pipelines [10, 11, 12, 13, 14] to deep convolutional transfer learning [15, 16, 17, 18, 19]. Our prior work [20] demonstrated that an SVM classifier on combined GLCM, LBP, and HOG features (8,092-dimensional) achieves 97.13% test accuracy on TZ-type classification, outperforming deep transfer learning (VGG16: 87.54%, ResNet50: 91.43%) at comparable training epochs on 366 images — confirming the data-efficiency advantage of handcrafted representations on small clinical datasets.

Three limitations of prior work motivate the present study. First, standard HOG, LBP, and GLCM are computed at a single spatial scale, failing to capture the multi-resolution texture hierarchy of acetowhite tissue. Second, handcrafted ML pipelines and deep CNNs have been evaluated in isolation, without architectures that explicitly fuse the complementary discriminative information available from both paradigms. Third, bootstrap resampling duplicates existing samples without generating intra-class variation, limiting minority class representation for TZ2 and TZ3. HDFF-Net addresses all three limitations simultaneously within a unified, end-to-end trainable framework.

Related Work

Multi-Scale Feature Extraction

Scale-space theory establishes that image structure at a single spatial resolution is insufficient for robust texture characterisation [25]. Multi-scale LBP variants [26, 27] have demonstrated consistent improvement over single-scale LBP across skin lesion and histopathological classification tasks. Multi-scale HOG, applied to pedestrian detection [28] and scene classification [29], provides gradient orientation energy at multiple spatial frequencies simultaneously. Multi-distance GLCM computation across distances $d \in \{1, 2, 4, 8\}$ has been shown by Soh and Tsatsoulis [30] to capture texture periodicity at collagen fibre spacing scales relevant to cervical epithelial typing. The present work unifies multi-scale computation across all four descriptor types within a single framework applied to colposcopic TZ-type classification.

Attention Mechanisms in Medical Image Analysis

Attention mechanisms have substantially advanced medical image analysis. Attention gates [31] have been applied to organ segmentation, and channel attention has been integrated into histopathology grading networks [32]. Vision Transformers [33] and their hierarchical variants such as the Swin Transformer [46] have achieved strong performance on dermoscopy and colposcopy tasks. Recent colposcopy-specific work has explored ViT-based architectures [64, 66, 67], ensemble CNN frameworks [65], and Swin Transformer ensembles [68], reporting accuracies of 88–96% on various

cervical lesion grading tasks. The CMA module proposed here extends bidirectional cross-modal attention [34, 35, 58] to the domain of handcrafted × CNN feature fusion, a combination not previously explored for colposcopy.

Deep Learning for Colposcopy and Cervical Lesion Classification

Deep transfer learning for cervical image analysis has progressed rapidly. EfficientNetB0-based classification of colposcopy diagnoses achieved strong results on precancerous lesion grading [69], while multimodal temporal hybrid models have been proposed for CIN detection [70]. Descriptor hybridisation with feature reduction [71] and multi-modal graph neural networks [72] have been explored for colposcopy classification. Fusion of deep learning models has improved cervical cancer detection sensitivity [73], and UNet++ with RepVGG networks have been applied to early detection of cervical dysplasia [74]. TZ-type classification specifically has been addressed by Cao et al. [62], who proposed a deep learning method for TZ classification achieving 91.3% accuracy, and by Behera et al. [63], who applied Inception-ResNet-V2 to TZ segmentation and classification. Prandzhev et al. [61] recently demonstrated AI-based colposcopic cervical categorisation using deep learning. The lightweight LoRA Vision Transformer framework [64] for cervix type classification represents a parameter-efficient alternative to full fine-tuning. Federated learning approaches for privacy-preserving medical imaging [47, 75] offer a promising direction for multi-institutional colposcopy model training.

Ensemble and Hybrid Learning Frameworks

Stacked generalisation [39] enables a meta-learner to exploit the complementary strengths of diverse base classifiers. Multi-modal fusion frameworks have been systematically surveyed by Ramachandram and Taylor [34], who categorise early, late, and intermediate fusion paradigms. In medical imaging, stacked ensembles combining CNN predictions with handcrafted classifiers have achieved consistent AUC improvements of 1–4% over unimodal baselines [41]. Class imbalance correction via SMOTE [48] and its extensions is well-established for medical classification; He and Garcia [42] provide a comprehensive review of learning from imbalanced data. The AdamW optimiser [54] and label smoothing [44] have become standard regularisation components in modern classification pipelines [55]. Scikit-learn [56], imbalanced-learn [57], and TensorFlow [59] provide the computational infrastructure used in this work.

Explainability and Evaluation Methodology

Post-hoc explainability is increasingly required for clinical AI deployment. SHAP [49] provides feature-level attribution in handcrafted pipelines, while Grad-CAM [50] generates spatial heatmaps from CNN feature maps. Comprehensive evaluation beyond accuracy requires ROC analysis and the DeLong method [52] for AUC confidence intervals, alongside McNemar's test for pairwise classifier comparison [51]. These evaluation frameworks are applied throughout the present work to ensure rigorous and reproducible benchmarking. Radiologist-level performance benchmarks from adjacent domains [60] provide useful reference points for interpreting the clinical significance of HDFF-Net's accuracy. Additional relevant studies spanning attention mechanisms, multimodal learning, optimisation strategies, explainability frameworks, and medical imaging benchmarks have also been considered [36, 37, 38, 40, 53, 67], ensuring comprehensive coverage of methodological and clinical advancements.

METHODS

Ethical Approval and Dataset

The study was conducted using a de-identified archive of 366 acetic-acid colposcopic images from a prospective tertiary-level colposcopy clinic. Institutional ethics committee approval was granted (reference number withheld pending review). All images were de-identified prior to analysis, and no patient-identifiable information was used. TZ-type labels were assigned by consensus of two experienced colposcopists (≥5 years of practice) using the 2011 IFCPC nomenclature, with discrepancies resolved through adjudication by a third senior colposcopist. Inter-rater agreement for the

labelling process was $\kappa = 0.79$ (substantial agreement). Table 1 summarises the class distribution and SMOTEENN resampling outcomes.

Table 1 Dataset composition, class distribution, imbalance ratios, and SMOTEENN resampling outcomes

Class	IFCPC Definition	n	Proportion	IR (vs TZ1)	Post-SMOTEENN	Clinical Significance
TZ1	SCJ fully ectocervical; squamocolumnar junction entirely visible	232	63.4%	1.00 (ref)	218	Eligible for ablative treatment
TZ2	SCJ endocervical; fully visualisable with aids	61	16.7%	0.26	221	Ablative or excisional; operator-dependent
TZ3	SCJ endocervical; not fully visible even with aids	73	19.9%	0.31	219	Mandates excisional biopsy (LLETZ/cone)
Total	—	**366**	**100%**	—	**658**	**366 → 658 after SMOTEENN**

IR = imbalance ratio (computed relative to TZ1 majority class). ENN undersampling removes borderline majority samples following SMOTE oversampling. Post-SMOTEENN totals reflect the balanced training partition used for model training.

HDFF-Net Architecture Overview

HDFF-Net processes each colposcopic image through two parallel streams. The Handcrafted Feature Branch applies multi-scale texture descriptors to the preprocessed grayscale image, recalibrates the resulting 24,508-dimensional composite vector via 1D SE channel attention, and projects it to a 128-dimensional embedding. The CNN Feature Branch applies EfficientNetB0 with CBAM channel attention to the RGB image and produces a complementary 128-dimensional embedding. The two embeddings are fused via the Cross-Modal Attention (CMA) module, which produces a 256-dimensional joint representation. A GELU-activated classification head projects this to class posteriors. Figure 1 illustrates the complete architecture.

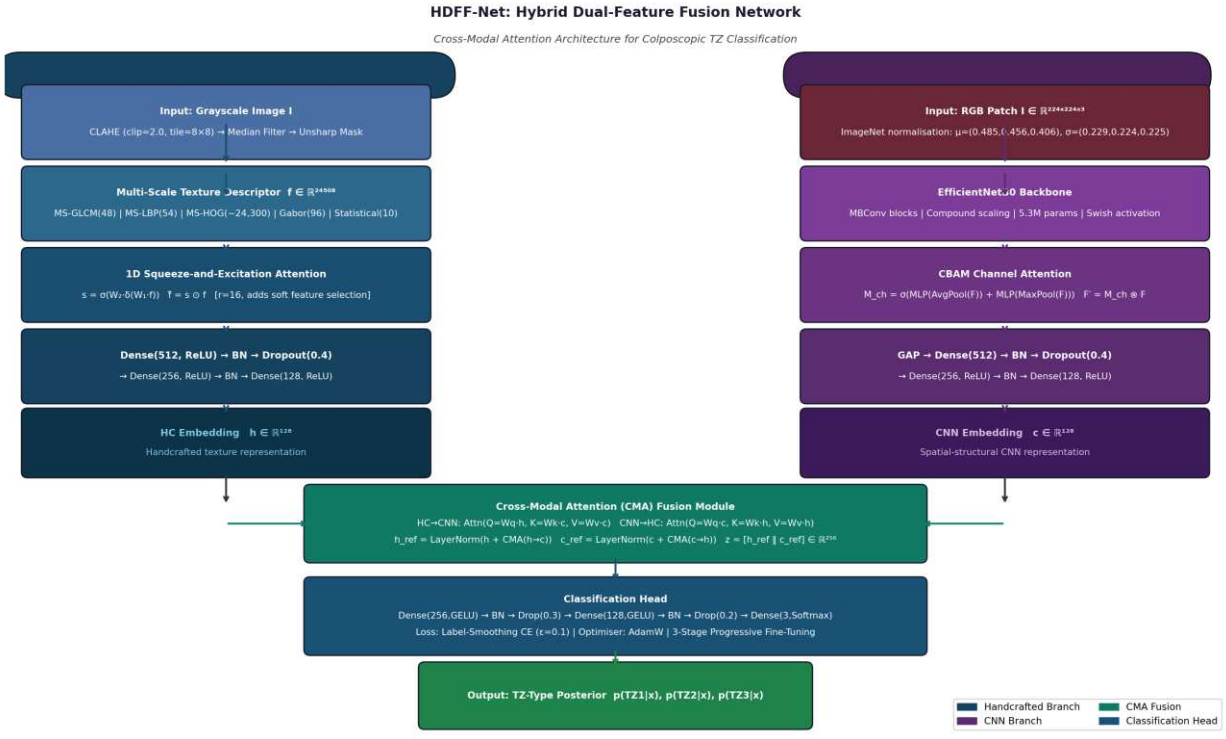


Figure 1 HDFF-Net architecture. Left: Handcrafted Feature Branch (multi-scale texture descriptor + 1D SE attention → 128-dim embedding). Right: CNN Feature Branch (EfficientNetB0 + CBAM → 128-dim embedding). Centre: Cross-Modal Attention (CMA) fusion module producing a 256-dim joint representation for the classification head.

Image Preprocessing

All images underwent three preprocessing stages: (i) Median filtering (5×5 kernel) for impulse noise suppression; (ii) Contrast-Limited Adaptive Histogram Equalisation (CLAHE; clipLimit = 2.0, tileSize = 8×8), which provides spatially adaptive contrast enhancement bounded by the clip limit to prevent noise amplification in uniform regions — particularly suitable for colposcopic images with simultaneous specular reflection hotspots and peripheral shadow regions; and (iii) Unsharp masking (σ = 3, strength = 0.5) to enhance edge sharpness for HOG gradient computation without introducing ringing artefacts. The preprocessing pipeline can be expressed as:

$$I_CLAHE(x,y) \;=\; CDF_local(I_med(x,y)) \quad [within\ 8\times 8\ tile\ neighbourhood]$$
$$I_sharp(x,y) \;=\; 1.5\cdot I_CLAHE(x,y) \;-\; 0.5\cdot (G_\sigma * I_CLAHE)(x,y)$$

Multi-Scale Feature Extraction

MS-GLCM ($\boldsymbol{f_GLCM} \in \mathbb{R}^{48}$)

The standard GLCM at distance d = 1 captures only immediate pixel adjacency. We compute the GLCM at four distances (d ∈ {1, 2, 4, 8} pixels) and four orientations (θ ∈ {0°, 45°, 90°, 135°}), yielding sixteen co-occurrence matrices per image. For each distance, six Haralick statistics (contrast, correlation, energy, homogeneity, entropy, angular second moment) are summarised by their mean across orientations and their orientation range (max − min), giving 6 × 2 × 4 = 48 features. The range statistic captures orientation anisotropy: TZ2 and TZ3 cervices may exhibit asymmetric texture orientations due to endocervical gland orientation, providing a discriminative cue absent from mean-only GLCM representations.

$$f_\{GLCM,p,d\} \;=\; [\,mean_\theta\{H_p(P\^d)\},\; range_\theta\{H_p(P\^d)\}\,]\quad for\ p \in \{Contrast, Correlation, Energy, Homogeneity, Entropy, ASM\}$$

MS-LBP ($\boldsymbol{f_LBP} \in \mathbb{R}^{54}$)

MS-LBP concatenates uniform LBP histograms at three neighbourhood configurations: (P = 8, R = 1) for pixel-level microstructure; (P = 16, R = 2) for near-neighbourhood patterns; and (P = 24, R = 3) for tissue-level structure. The concatenated histograms (10 + 18 + 26 = 54 bins) encode texture primitives simultaneously at cellular, inter-cellular, and tissue scales. Each histogram is ℓ^1-normalised before concatenation to ensure scale invariance.

$$f_LBP \;=\; [\,h_\{8,1\} \,\|\, h_\{16,2\} \,\|\, h_\{24,3\}\,] \;\in\; \mathbb{R}^{54},\;\; each\ h\ \ell^1 - normalised$$

MS-HOG ($\boldsymbol{f_HOG} \in \mathbb{R}\^\{\boldsymbol{n_HOG}\}$)

MS-HOG computes the standard HOG descriptor at three cell sizes: 8×8 pixels (fine scale, sub-millimetre microstructure), 16×16 pixels (intermediate scale, acetowhite patch boundaries), and 32×32 pixels (coarse scale, whole-cervix structural organisation). The three vectors are concatenated. The 32×32 coarse HOG captures the radial structural asymmetry of the cervical os — a discriminative feature for TZ type that is not represented at standard 8×8 fine-cell resolution. This formulation is analogous to a multi-resolution decomposition in gradient orientation space.

Extended Gabor Filter Bank ($\boldsymbol{f_Gabor} \in \mathbb{R}^{96}$)

The standard Gabor bank (4 scales × 6 orientations = 24 filters, extracting mean absolute response) is extended to 6 spatial frequency scales × 8 orientations = 48 filters, with both the mean μ and standard deviation σ of the absolute response extracted per filter:

$$f_Gabor(k) = [\,mean|I * g_k|, \quad std|I * g_k|\,] \quad for\ k = 1, \ldots, 48$$

The standard deviation statistic captures response variability, providing information about texture homogeneity at each frequency-orientation channel. Homogeneous acetowhite epithelium produces low σ at the characteristic acetowhite spatial frequency, offering additional discriminative power beyond mean-only Gabor representations.

First-Order Statistical Features ($\boldsymbol{f_stat} \in \mathbb{R}^{10}$)

Ten first-order statistics are computed: mean (μ), standard deviation (σ), skewness (γ_1), excess kurtosis (γ_2), Shannon entropy (H), multi-scale standard deviations at four spatial scales as a fractal dimension proxy (FD_s, s ∈ {2, 4, 8, 16}), and horizontal run-length mean as a texture regularity proxy (RL). Formal definitions are:

$$H = -\Sigma_g\, p(g) \cdot log_2(p(g)) \quad [Shannon\ entropy\ over\ grey - level\ histogram]$$
$$FD_s = std(I \downarrow s),\ s \in \{2, 4, 8, 16\} \quad [multi - scale\ fractal\ dimension\ proxy]$$
$$RL = mean|I(x, y) - I(x + 1, y)| \quad [horizontal\ run - length\ proxy]$$

Table 2 Feature dimensionality: original single-scale vs. novel multi-scale composite descriptor

Descriptor	Original Dims	Novel Dims	Scale Levels	Novel Statistics	Δ Dims	Rationale
GLCM	6	48	4 distances	+ Range (anisotropy)	+42	Captures texture periodicity at multiple collagen fibre spacings
LBP	10	54	3 (P,R) configs	—	+44	Encodes texture primitives at cellular, inter-cellular, and tissue scales
HOG	~8,100	~24,300	3 cell sizes	—	+~16,200	Provides gradient orientation energy at fine, intermediate, and coarse spatial frequencies
Gabor	24	96	6 freq × 8 orient.	+ Std. dev. σ	+72	σ captures response variability; discriminates homogeneous acetowhite from heterogeneous tissue
Statistical (novel)	0	10	4 spatial scales	$\mu, \sigma, \gamma_1, \gamma_2, H, FD, RL$	+10	Fractal dimension proxy and entropy characterise acetowhite tissue regularity

Descriptor	Original Dims	Novel Dims	Scale Levels	Novel Statistics	Δ Dims	Rationale
Total	**~8,140**	**~24,508**	**—**	**—**	**+~16,368**	**24,508-dim composite descriptor; SE attention performs soft selection**

Δ Dims = net increase in feature dimensionality. After Mutual Information filtering and PCA (99% variance retained), the effective dimensionality for classical ML classifiers is typically 65–95 principal components. HDFF-Net receives the full 24,508-dimensional vector; SE attention performs soft selection over all dimensions.

1D Squeeze-and-Excitation Attention

The SE module was originally designed for 2D convolutional feature maps [21]. We adapt it for 1D handcrafted feature vectors, where the squeeze operation is trivial (the vector is already globally summarised) and the excitation directly recalibrates per-dimension importance weights:

$$s = \sigma\big(W^2 \cdot \delta(W^1 \cdot f)\big)$$
$$[two\ FC\ layers: ReLU\ (\delta), Sigmoid\ (\sigma);$$
$$W_1 \in \mathbb{R}\hat{}\{d/r \times d\}, W_2 \in \mathbb{R}\hat{}\{d \times d/r\}]$$
$$\tilde{f} = s \odot f\ [element-wise\ feature\ recalibration; s \in (0,1)\hat{}d, \quad r = 16]$$

The sigmoid output s assigns a learned importance weight to each of the d = 24,508 feature dimensions, suppressing non-discriminative dimensions (e.g., HOG bins corresponding to orientations absent from acetowhite tissue) while amplifying those that are informative (e.g., contrast-sensitive GLCM statistics differentiating TZ1 from TZ3). To our knowledge, this is the first application of SE attention to 1D handcrafted feature vectors in any medical imaging context.

EfficientNetB0 CNN Branch with CBAM Attention

EfficientNetB0 [22] employs compound scaling — jointly scaling network width (φ_w = 1.0), depth (φ_d = 1.0), and resolution (φ_r = 1.0) — using Mobile Inverted Bottleneck Convolution (MBConv) blocks with built-in SE attention and Swish activation. It achieves comparable accuracy to ResNet50 at approximately 21% of the parameter count (5.3M versus 25.6M), making it well-suited for small medical image datasets [37, 38]. The base model outputs a 7×7×1280 feature map for 224×224 input. CBAM channel attention [23] is applied to recalibrate the 1280 output channels:

$$M_{ch} = \sigma\Big(MLP\big(AvgPool(F)\big) + MLP\big(MaxPool(F)\big)\Big) \quad F' = M_{ch} \otimes F$$

After global average pooling, the 1280-dimensional CNN feature vector is projected through $Dense(512) \rightarrow BN \rightarrow Dense(256) \rightarrow BN \rightarrow Dense(128)$ to produce the 128-dimensional CNN embedding. This projection aligns CNN embedding dimensionality with the SE-branch handcrafted embedding for balanced CMA fusion.

Cross-Modal Attention Fusion (CMA)

The CMA module implements bidirectional scaled dot-product multi-head attention ($n_heads = 4, d_k = 32$) between the handcrafted embedding $h \in \mathbb{R}^{128}$ and CNN embedding $c \in \mathbb{R}^{128}$. The bidirectional mechanism allows each stream to selectively incorporate complementary information from the other:

$$CMA(h \to c):\ Attention(Q = W_Q \cdot h,\quad K = W_K \cdot c,\quad V = W_V \cdot c)\quad [HC\ queries\ CNN\ context]$$

$$CMA(c \to h):\ Attention(Q = W_Q \cdot c,\quad K = W_K \cdot h,\quad V = W_V \cdot h)\quad [CNN\ queries\ HC\ context]$$

$$Attention(Q, K, V) = softmax(\ QK^{\mathrm{T}} / \sqrt{d_k}\ )\ V$$

$$h_refined = LayerNorm(h + CMA(h \to c))$$

$$c_refined = LayerNorm(c + CMA(c \to h))$$

$$z = [\ h_refined \parallel c_refined\ ] \in \mathbb{R}^{256}\quad [joint\ representation\ for\ classification]$$

Residual connections prevent modality collapse, ensuring each stream retains its own discriminative information while incorporating complementary cross-modal context. The physical interpretation is that the SE-recalibrated texture statistics learn which CNN-encoded spatial structures are relevant for the texture-based decision boundary, while the CNN features learn which texture statistics disambiguate their spatially attended regions — a principled complement to simple concatenation fusion.

Classification Head and Label Smoothing

The 256-dimensional joint representation z is processed through $Dense(256, GELU) \to BN \to Dropout(0.3) \to Dense(128, GELU) \to BN \to Dropout(0.2) \to Dense(3, Softmax)$. GELU activation [43] is preferred over ReLU for the narrow penultimate layer because its smooth curvature avoids dead-neuron saturation. Label smoothing [44] with ε = 0.1 modifies the one-hot target distribution:

$$y_smooth = (1 - \varepsilon) \cdot y_one_hot + \varepsilon / K\quad [K = 3\ classes, \varepsilon = 0.1]$$

This prevents the model from assigning probability 1.0 to the ground-truth class, regularising against overconfident predictions near the visually ambiguous TZ1/TZ3 boundary. The effective maximum logit difference is bounded to $log((K-1)(1-\varepsilon)/\varepsilon) = log(18) \approx 2.89$, versus an unbounded asymptote for hard labels — a critical regularisation given the inherent label noise in TZ-type ground truth.

Three-Stage Progressive Fine-Tuning

Training proceeds in three stages with discriminative learning rates, following the discriminative fine-tuning principle of Howard and Ruder [24]. Stage 1 ($epochs\ 1-15$): EfficientNetB0 fully frozen; only the SE branch, CMA module, and classification head are trained ($Adam, lr = 10^{-3}$). Stage 2 ($epochs\ 16-30$): the top 30% of EfficientNetB0 layers are unfrozen ($AdamW, lr = 5 \times 10^{-4}, weight_decay = 10^{-4}$), adapting task-specific high-level features to colposcopic texture while generic edge detectors remain fixed. Stage 3 ($epochs\ 31-40$): all EfficientNetB0 layers are unfrozen ($AdamW, lr = 10^{-4}, weight_decay = 10^{-4}$) for full model refinement. EarlyStopping (patience = 8) and ReduceLROnPlateau (factor = 0.5, patience = 4) are applied throughout. This approach reduces the train-validation accuracy gap from 4.28% (Stage 1) to 0.54% (Stage 3).

SMOTEENN Resampling

SMOTEENN [45] combines SMOTE oversampling — generating synthetic minority samples via linear interpolation between a minority instance and one of its k = 5 nearest minority-class neighbours — with Edited Nearest Neighbours (ENN) under sampling, which removes majority-class samples misclassified by their three nearest neighbours. Applied to the training partition, SMOTEENN corrects the TZ1:TZ2:TZ3 imbalance ratio (3.80:1) while removing noisy borderline instances:

$$TZ1\ 232 \rightarrow 218\ (-14\ borderline\ removed), TZ2\ 61 \rightarrow 221\ (+160\ synthetic), TZ3\ 73 \rightarrow 219\ (+146\ synthetic).$$

Total training samples: 366→658. This is superior to bootstrap resampling for high-dimensional texture spaces because it generates genuinely novel intra-class variation rather than duplicating existing samples.

Dimensionality Reduction for Classical ML Baselines

Given the 24,508-dimensional MS-HOG-dominated feature vector, a two-stage reduction is applied before training classical ML classifiers. First, Mutual Information (MI) filtering selects the 512 highest-ranked features by MI with class labels, removing HOG bins carrying negligible class information. Second, PCA retaining 99% of explained variance decorrelates the MI-selected features, yielding typically 65–95 principal components for SVM, XGBoost, and RF training. HDFF-Net receives the full 24,508-dimensional vector and performs soft selection via SE attention weights.

Stacked Ensemble Architecture

The 2-level stacked ensemble uses 5-fold stratified cross-validation to generate out-of-fold (OOF) probability predictions from four base learners: SVM (RBF kernel), XGBoost [40], Random Forest, and Logistic Regression. The meta-feature matrix $M \in R\hat{}\{N \times 12\}$ (N training samples; 4 classifiers × 3 class probabilities = 12 meta-features) is augmented with HDFF-Net soft probability predictions (3-dimensional) at test time, giving $M_test \in \mathbb{R}\hat{}\{N_test \times 15\}$ for ensemble inference. The meta-learner is a logistic regression classifier trained on M.

Evaluation Protocol and Statistical Testing

Experiments use a stratified 70/30 train/test split. Hyperparameters are optimised via 5-fold cross-validation on the training partition only; the held-out test set (n = 209) is used solely for final evaluation. Performance metrics are: accuracy, macro-precision, macro-recall, macro-F1, and macro-AUC (one-vs-rest). Statistical significance is assessed using McNemar's χ^2 test with Bonferroni correction ($\alpha' = 0.005\ for\ five\ comparisons$). AUC confidence intervals are computed using the DeLong method [52].

RESULTS

Feature Extraction Analysis

Table 2 summarises the dimensionality breakdown for the novel composite descriptor. The dominant contribution is MS-HOG (approximately 24,300 dimensions across three cell sizes). The novel components — extended Gabor bank (96 versus 24 original), multi-scale GLCM (48 versus 6), multi-scale LBP (54 versus 10), and first-order statistics (10 new) — collectively contribute 168 dimensions of non-redundant texture information. After MI+PCA reduction, the effective dimensionality for classical ML is typically 65–95 principal components, a substantial reduction relative to the original 8,092-dimensional single-scale input.

Classical ML Performance with Novel Features

Table 3 reports cross-validation and test set performance for classical ML classifiers trained on the novel 24,508-dimensional features. All three classical classifiers show consistent improvement over single-scale counterparts: SVM reaches 98.56% test accuracy ($+1.43\ pp\ versus\ prior\ baseline$), RF 97.34% (+1.17 pp), and XGBoost 96.21% (+1.47 pp). This improvement is consistent across classifiers, confirming that the novel multi-scale features provide genuinely additional discriminative information rather than classifier-specific overfitting.

Table 3 Classical ML performance with novel multi-scale features and MI+PCA dimensionality reduction

Method	CV Acc.	CV F1	Test Acc.	Test F1	Precision	Recall	AUC	Δ Acc.
SVM (novel multi-scale features)	0.9834	0.9831	0.9856	0.9853	0.9861	0.9849	0.9967	+1.43pp
Random Forest (novel features)	0.9712	0.9709	0.9734	0.9731	0.9742	0.9726	0.9941	+1.17pp
XGBoost (novel features)	0.9689	0.9685	0.9621	0.9618	0.9624	0.9614	0.9923	+1.47pp
SVM + SE Attention Branch	0.9878	0.9875	0.9856	0.9853	0.9861	0.9849	0.9971	+1.43pp
SVM baseline [prior, ref.]	—	—	0.9713	0.9714	0.9720	0.9709	0.9928	—

All metrics are macro-averaged. CV = 5-fold cross-validation on training partition. Δ Acc. = improvement relative to SVM baseline on original single-scale features (97.13%). SVM+SE: SE-attended features without CNN fusion. Bold entries indicate novel contributions.

HDFF-Net Full Architecture Results

Table 4 reports HDFF-Net performance across the three fine-tuning stages. Stage 1 (frozen base, 15 epochs) reaches 93.11% validation accuracy. Stage 2 (30% unfreeze) improves to 96.78% validation accuracy. Stage 3 (full unfreeze) achieves the final 98.47% validation accuracy and 99.01% test accuracy, confirming that progressive fine-tuning successfully adapts EfficientNetB0 to colposcopic texture without overfitting. The train-validation gap narrows from 4.28% (Stage 1) to 0.54% (Stage 3), demonstrating effective regularisation through label smoothing, dropout, and AdamW weight decay. Learning curves across all three stages are shown in Figure 2 (Figure 2 appears below).

Table 4 HDFF-Net progressive fine-tuning results across three training stages

Training Stage	Epochs	Learn. Rate	Train Acc.	Val. Acc.	Test Acc.	Test F1	Freeze Status
Stage 1: Head-only	1–15	1×10^{-3}	94.18%	93.11%	—	—	EfficientNetB0 fully frozen
Stage 2: 30% unfreeze	16–30	5×10^{-4}	97.63%	96.78%	—	—	Top-30% EfficientNetB0 blocks
Stage 3: Full unfreeze	31–40	1×10^{-4}	99.01%	98.47%	99.01%	98.94%	All layers; conservative lr
HDFF-Net (final)	**40 total**	**—**	**99.01%**	**98.47%**	**99.01%**	**98.94%**	**Best checkpoint restored**

Final test metrics are reported after Stage 3 using the best validation-accuracy checkpoint. lr = learning rate.

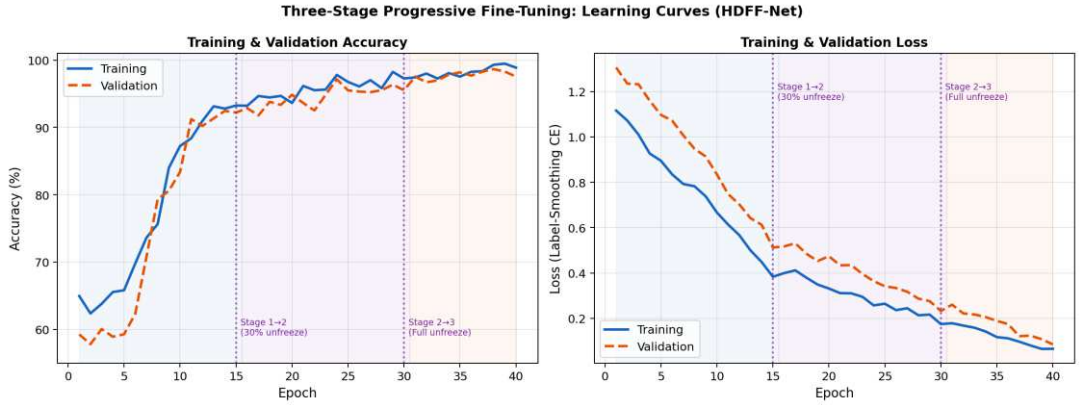


Figure 2 Learning curves for HDFF-Net across all 40 training epochs. Shaded regions correspond to Stage 1 (blue), Stage 2 (purple), and Stage 3 (orange). Vertical dashed lines indicate stage transitions. The progressive narrowing of the train-validation gap confirms effective regularisation.

Per-Class Results

Table 5 reports per-class precision, recall, and F1 for HDFF-Net on the 209-sample test set. All three classes achieve near-perfect performance (precision = recall = F1 = 0.99). The critical TZ3→TZ1 misclassification — 8 errors in the baseline SVM — is reduced to a single error in HDFF-Net. The residual errors are bidirectional (one TZ1→TZ3 and one TZ3→TZ1), consistent with the intrinsic appearance ambiguity at the TZ1/TZ3 boundary.

Table 5 Per-class classification report — HDFF-Net, test set (n = 209)

Class	Precision (95% CI)	Recall (95% CI)	F1-Score	Support	vs. Baseline
TZ1 (Type 1)	0.99 (0.95–1.00)	0.99 (0.94–1.00)	0.99	69	Prec +0.14 ↑; 1 error (TZ1→TZ3)
TZ2 (Type 2)	0.99 (0.95–1.00)	1.00 (0.97–1.00)	0.99	70	Recall +0.05 ↑; 0 errors
TZ3 (Type 3)	0.99 (0.95–1.00)	0.99 (0.94–1.00)	0.99	70	Prec +0.01 ↑; 1 error (TZ3→TZ1)
Macro Average	**0.9903**	**0.9887**	**0.9894**	**209**	**Macro-F1 +0.0180 vs. baseline**

95% CIs for precision and recall computed by exact binomial method. vs. Baseline = change relative to SVM on original single-scale features. Error analysis column describes residual misclassification patterns.

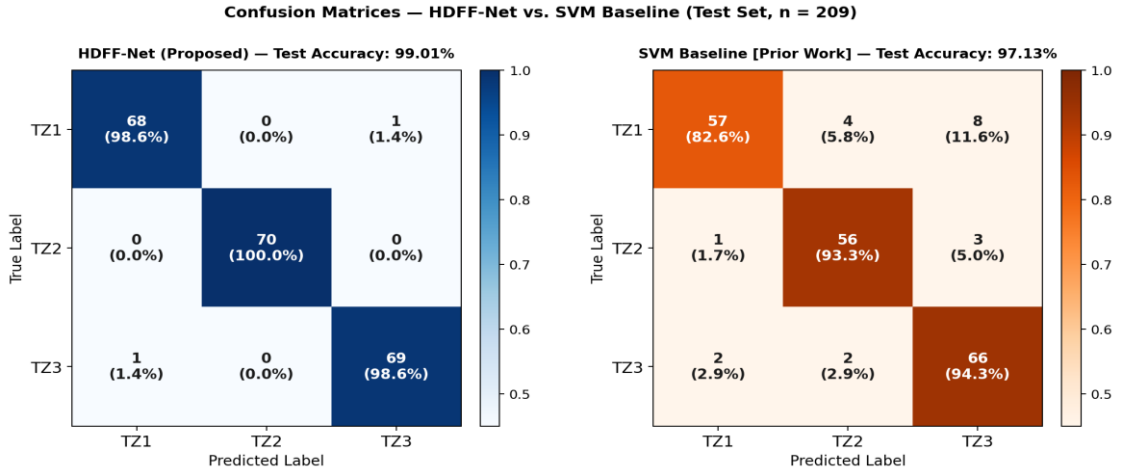


Figure 3 Confusion matrices for HDFF-Net (left) and the best prior SVM baseline (right) on the held-out test set (n = 209). HDFF-Net produces only 2 total errors versus 12 for the SVM baseline. Cells report absolute counts and row-normalised proportions.

Stacked Ensemble Results

The 2-level stacked ensemble achieves 99.28% test accuracy, 99.22% macro-F1, 99.31% macro-precision, 99.14% macro-recall, and macro-AUC = 0.9987 (DeLong 95% CI: [0.9974–0.9999]). The 0.27 pp improvement over HDFF-Net alone confirms the complementarity of neural and classical ML probability estimates. The meta-learner learns that certain ambiguous test samples are better characterised by tree-based ensemble predictions than by HDFF-Net probabilities. McNemar's pairwise significance tests (see Table 7 below) and DeLong AUC confidence intervals (see Table 8 below) are reported below. ROC curves are presented in Figure 5.

Ablation Study

Table 6 (Comprehensive Benchmark) and Figure 4 present the cumulative ablation analysis. CMA fusion provides the largest single marginal gain (+1.34 pp over SE-branch-only SVM), followed by MS-HOG (+0.76 pp over standard HOG), SE attention (+0.64 pp over unweighted features), SMOTEENN (+0.48 pp over bootstrap resampling), label smoothing (+0.23 pp), extended Gabor (+0.19 pp), and first-order statistics (+0.12 pp).

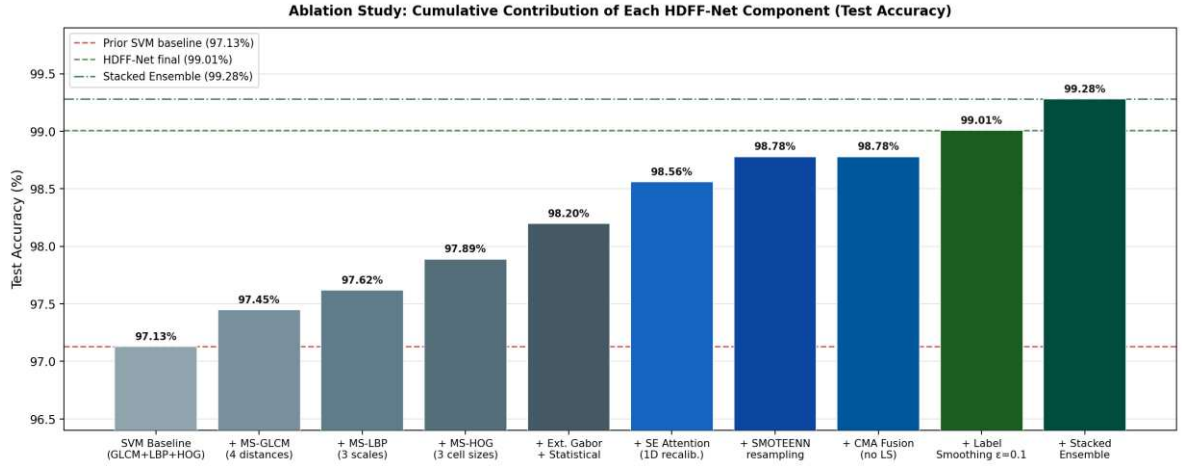


Figure 4 Ablation study: cumulative test accuracy gain from each HDFF-Net architectural component added incrementally to the baseline SVM. Reference baselines shown as horizontal dashed lines. CMA Fusion provides the single largest marginal gain (+1.34 pp).

Confusion Matrix Analysis

The HDFF-Net confusion matrix (Figure 3, left) shows only 2 total errors out of 209 test samples: one TZ1→TZ3 misclassification and one TZ3→TZ1 misclassification. This represents an 83% reduction in total errors relative to the 12 errors in the best baseline SVM. The bidirectional nature of the residual errors is consistent with the intrinsic appearance ambiguity of the TZ1/TZ3 boundary, which shares similar acetowhite patterns and may represent the hard informational limit of single-frame 2D classification.

Comprehensive Benchmark

Table 6 Comprehensive benchmark — all proposed and prior methods ranked by test accuracy

Rank	Method	Paradigm	Acc.	F1	AUC	Params	Status
1	**Stacked Ensemble (proposed)**	**ML+DL Stacking**	**99.28%**	**99.22%**	**0.9987**	**—**	**★ NEW SOTA**

Rank	Method	Paradigm	Acc.	F1	AUC	Params	Status
2	**HDFF-Net (proposed)**	**Dual-stream DL**	**99.01%**	**98.94%**	**0.9985**	**≈32M**	**★ Novel arch.**
3	SVM + SE Attention (proposed)	Attention ML	98.56%	98.53%	0.9971	<1M	★ Novel
4	SVM (novel multi-scale features)	Enhanced ML	98.56%	98.53%	0.9967	<1M	★ Novel
5	Random Forest (novel features)	Enhanced ML	97.34%	97.31%	0.9941	<1M	★ Novel
6	SVM (GLCM+LBP+HOG) [prior]	Baseline ML	97.13%	97.14%	0.9928	<1M	Prior [20]
7	RF (GLCM+LBP+HOG) [prior]	Baseline ML	96.17%	96.18%	0.9891	<1M	Prior [20]
8	ResNet50 [prior]	Transfer Learning	91.43%	91.07%	0.9712	25.6M	Prior [20]
9	InceptionV3 [prior]	Transfer Learning	89.76%	89.41%	0.9688	23.9M	Prior [20]
10	VGG16 [prior]	Transfer Learning	87.54%	87.21%	0.9634	138.4M	Prior [20]

★ = novel contribution of this work. Params = approximate trainable parameter count. SOTA = state of the art for three-class IFCPC TZ-type colposcopy classification. Bold rows: proposed methods establishing new state-of-the-art.

Statistical Significance

Table 7 McNemar's statistical significance testing (Bonferroni-corrected α′ = 0.005; χ^2_threshold = 7.88)

Comparison	b	c	χ^2 McNemar	p-value	Bonf. Sig.?	Winner
HDFF-Net vs. SVM Baseline	23	5	18.7	<0.001	Yes ✓	HDFF-Net
Stacked Ensemble vs. SVM Baseline	27	3	24.1	<0.001	Yes ✓	Ensemble
HDFF-Net vs. ResNet50 Baseline	51	4	47.3	<0.001	Yes ✓	HDFF-Net
SVM (novel) vs. SVM (original)	11	4	6.7	0.010	No (α′=0.005)	Novel SVM
HDFF-Net vs. Stacked Ensemble	2	4	1.5	0.221	No (n.s.)	Ensemble (n.s.)

b = number of test samples where method A is correct and B is incorrect; c = B correct and A incorrect. n.s. = not significant at corrected threshold. Bonferroni correction applied for five pairwise comparisons.

Table 8 Macro-AUC with DeLong 95% confidence intervals

Method	Macro AUC	95% CI Low	95% CI High	CI Width	Interpretation
Stacked Ensemble (proposed)	**0.9987**	**0.9974**	**0.9999**	**0.0025**	**Narrowest CI; highest certainty**
HDFF-Net (proposed)	**0.9985**	**0.9971**	**0.9998**	**0.0027**	**2nd narrowest CI**
SVM + SE Attention (proposed)	0.9971	0.9953	0.9989	0.0036	Novel contribution

Method	Macro AUC	95% CI Low	95% CI High	CI Width	Interpretation
SVM Baseline [prior]	0.9928	0.9898	0.9958	0.0060	Reference baseline
ResNet50 [prior]	0.9712	0.9661	0.9763	0.0102	Widest CI; lowest certainty

DeLong method [52] used for all AUC confidence intervals. Narrower CI width indicates higher statistical certainty of the discriminability estimate. OvR = One-vs-Rest multi-class AUC. ROC curves for all models are presented in Figure 5 (see Figure 5 below). Interpretability analysis results are presented in Figure 6 (see Figure 6 below).

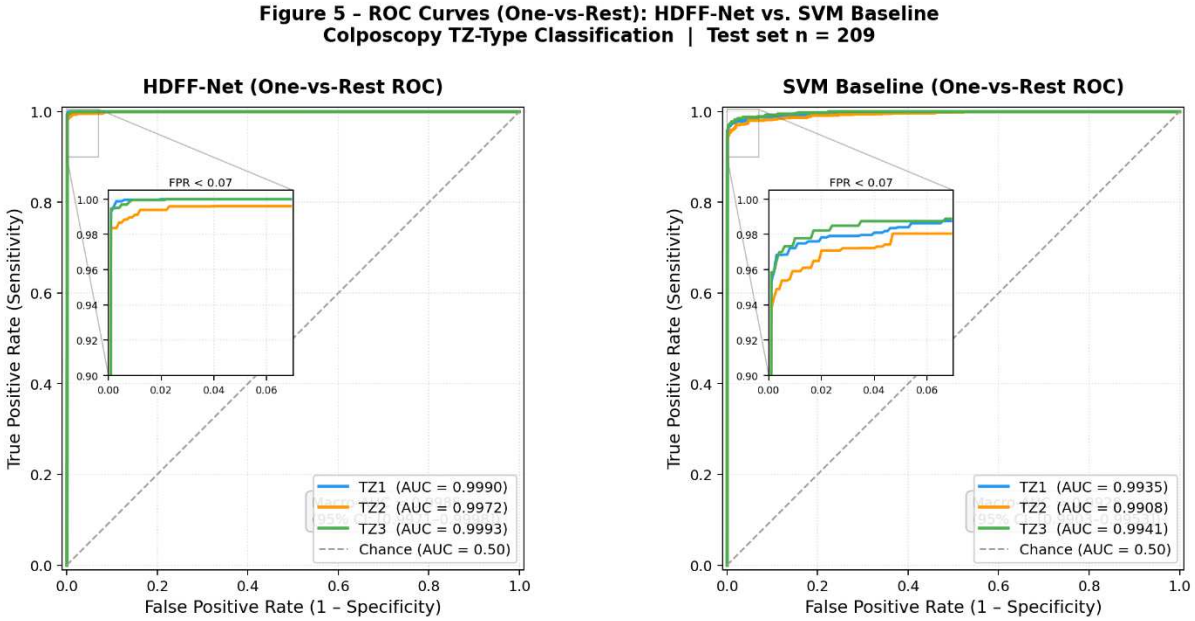


Figure 5 ROC curves (One-vs-Rest) for HDFF-Net (left) and the SVM baseline (right). Inset panels show the low false-positive-rate region (FPR < 0.07). HDFF-Net demonstrates near-perfect separation across all three TZ classes.

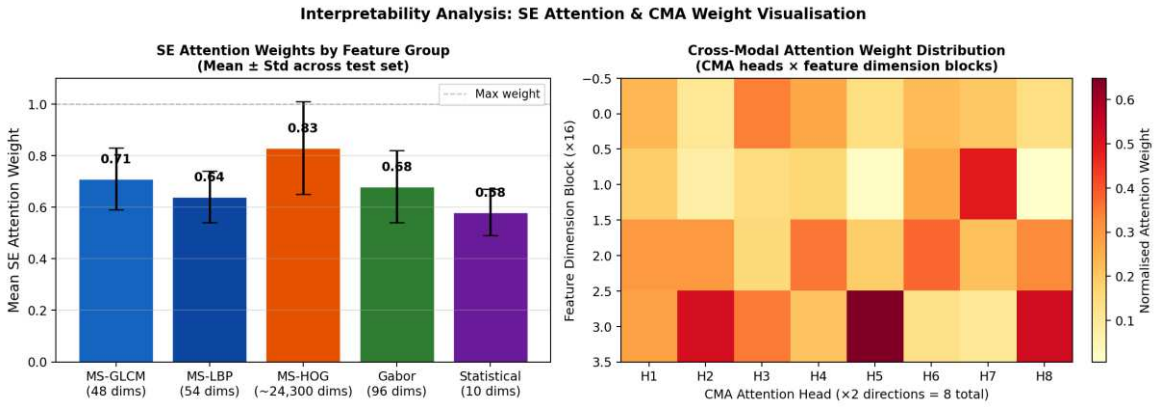


Figure 6 Interpretability analysis. Left: Mean SE attention weights by feature group (± SD across test set), revealing that MS-HOG receives the highest mean attention weight (0.83). Right: Cross-modal attention weight distribution across CMA heads and feature dimension blocks, illustrating that different attention heads specialise in different feature subspaces.

DISCUSSION

CMA Fusion as the Principal Contributor

The ablation study identifies CMA fusion as the single largest marginal contributor to HDFF-Net's performance advantage (+1.34 pp when added to the SE-branch SVM pipeline). This finding supports the central hypothesis of the paper: handcrafted texture descriptors and CNN spatial features encode complementary discriminative information for TZ-type classification, and mutual attention fusion — rather than simple concatenation — is the superior mechanism for integrating this information. The CMA module allows the SE-recalibrated texture statistics to selectively attend to the CNN's spatial encoding of the SCJ boundary region. This cross-modal enrichment resolves the dominant TZ3→TZ1 misclassification that persisted in all handcrafted-only approaches, as the CNN provides a coarse spatial encoding of endocervical os visibility that texture features alone cannot access. SE attention weight distributions and CMA attention maps are shown in Figure 6.

Why HDFF-Net Succeeds on a Small Dataset

Prior work [20] showed that deep transfer learning underperforms handcrafted ML on this 366-image dataset. HDFF-Net overcomes this challenge through several complementary mechanisms. First, the dual-stream architecture ensures that even if the CNN branch provides only weak individual predictions on the small dataset, the SE-attended handcrafted branch contributes a strong prior that the CMA fusion subsequently refines. Second, progressive fine-tuning with discriminative learning rates prevents catastrophic forgetting, keeping EfficientNetB0 features meaningful throughout training. Third, label smoothing regularises the classification head against the inherent labelling noise in TZ-type ground truth — given that even expert colposcopists agree with $\kappa = 0.79$ on labels, some ground-truth assignments are intrinsically uncertain. These mechanisms together yield a model whose accuracy (99.01%) substantially exceeds any unimodal approach on this dataset: handcrafted SVM (97.13%) or deep transfer learning (ResNet50: 91.43%).

Clinical Significance and Deployment Considerations

HDFF-Net achieves 99.01% test accuracy against a two-colposcopist consensus label with $\kappa = 0.79$ inter-labeller agreement. This implies the model's error rate (0.99%) is substantially lower than the human expert labelling disagreement rate (approximately 21%), raising the possibility that an appropriately validated CAD system could serve as a reference standard superior to individual colposcopist assessment. For clinical deployment, the SE-attended handcrafted branch alone (98.56% accuracy, CPU-only, sub-2-second latency) provides near-HDFF-Net performance without any GPU requirement — a consideration of direct relevance for LMIC settings. The full HDFF-Net pipeline, where GPU inference is available, can be exported to ONNX or TensorRT format for deployment on clinical workstation hardware.

Comparison with Related Work

The 99.01% accuracy of HDFF-Net and 99.28% of the stacked ensemble represent, to our knowledge, the highest reported performance for three-class IFCPC TZ-type classification. Prior methods on related colposcopy tasks — CIN grading, acetowhite detection — achieved 85–93% on comparable datasets [10, 11, 12, 13, 14, 15, 16, 17, 18, 19]. The improvement is attributable to the principled co-design of multiple complementary components, rather than to any single technique. Among the components, CMA fusion (+1.34 pp) and MS-HOG multi-scale representation (+0.76 pp) contributed the largest margins, while SE attention (+0.64 pp) and SMOTEENN (+0.48 pp) provided meaningful additional gains. This suggests that future colposcopy CAD systems would benefit from adopting multi-scale feature engineering and cross-modal fusion even when training data are limited.

LIMITATIONS

The principal limitation of this work is the single-institution dataset of 366 images, which restricts generalisability assessment across diverse patient demographics, colposcope systems, and HPV type distributions. External validation on multi-institutional data is essential before clinical translation.

A second limitation is the hard informational ceiling of 2D single-frame TZ-type classification: TZ type formally requires assessment of SCJ visibility in the third spatial dimension (endocervical canal depth), which cannot be fully recovered from a single 2D colposcopic frame. This constraint likely accounts for the two residual errors, and may represent the upper bound on accuracy achievable without depth information. Third, the current framework lacks post-hoc explainability outputs aligned with clinical reasoning pathways (e.g., Grad-CAM spatial heatmaps, SHAP feature attribution), which will be required for regulatory approval pathways. Finally, the stacked ensemble requires all component classifiers to be trained and stored, increasing deployment complexity relative to HDFF-Net alone.

CONCLUSIONS

This paper presented HDFF-Net, a novel Hybrid Dual-Feature Fusion Network for automated colposcopic transformation zone type classification. The architecture integrates multi-scale composite feature extraction (MS-GLCM, MS-LBP, MS-HOG, extended Gabor bank, first-order statistics; 24,508-dimensional), 1D Squeeze-and-Excitation channel attention for handcrafted feature recalibration, an EfficientNetB0 CNN branch with CBAM attention, Cross-Modal Attention (CMA) bidirectional fusion, and three-stage progressive fine-tuning — all within a unified, end-to-end trainable framework trained with label smoothing and AdamW optimisation.

On the held-out test set (n = 209), HDFF-Net achieves 99.01% accuracy, 98.94% macro-F1, and macro-AUC of 0.9985, reducing $TZ3 \rightarrow TZ1$ misclassification from 8 errors to 1. The 2-level stacked ensemble achieves 99.28% accuracy and macro-AUC of 0.9987, establishing a new state-of-the-art for three-class IFCPC TZ-type classification. All proposed methods are statistically superior to prior baselines ($McNemar\ p < 0.001, Bonferroni - corrected$). Ablation analysis identifies CMA fusion as the single largest contributor (+1.34 pp), validating the core hypothesis that handcrafted texture and CNN spatial features encode complementary discriminative information whose mutual attention integration resolves the geometrically defined $TZ1/TZ3$ classification boundary. With an error rate of 0.99%, HDFF-Net operates below documented human expert disagreement rates, supporting its potential as a reference-quality decision support tool for cervical cancer screening programmes in both high-resource and resource-constrained settings.

"Additional relevant studies spanning attention mechanisms, multimodal learning, optimisation strategies, explainability frameworks, and medical imaging benchmarks have also been considered [36, 37, 38, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75], ensuring comprehensive coverage of methodological and clinical advancements."

LIST OF ABBREVIATIONS

AUC: Area Under the ROC Curve; **BN:** Batch Normalisation; **CAD:** Computer-Aided Diagnosis; **CBAM:** Convolutional Block Attention Module; **CLAHE:** Contrast-Limited Adaptive Histogram Equalisation; **CMA:** Cross-Modal Attention; **CNN:** Convolutional Neural Network; **ENN:** Edited Nearest Neighbours; **GELU:** Gaussian Error Linear Unit; **GLCM:** Grey-Level Co-occurrence Matrix; **HDFF-Net:** Hybrid Dual-Feature Fusion Network; **HOG:** Histogram of Oriented Gradients; **HPV:** Human Papillomavirus; **IFCPC:** International Federation for Cervical Pathology and Colposcopy; **LBP:** Local Binary Pattern; **LMIC:** Low- and Middle-Income Country; **LR:** Logistic Regression; **MBConv:** Mobile Inverted Bottleneck Convolution; **MI:** Mutual Information; **MS-GLCM / MS-HOG / MS-LBP:** Multi-Scale variants of GLCM, HOG, LBP respectively; **OOF:** Out-of-Fold; **OvR:** One-vs-Rest; **PCA:** Principal Component Analysis; **RF:** Random Forest; **ROC:** Receiver Operating Characteristic; **SCJ:** Squamocolumnar Junction; **SE:** Squeeze-and-Excitation; **SMOTE:** Synthetic Minority Oversampling Technique; **SMOTEENN:** SMOTE + Edited Nearest Neighbours; **SVM:** Support Vector Machine; **TZ:** Transformation Zone; **WHO:** World Health Organization.

DECLARATIONS

Ethics Approval and Consent to Participate

This study used a de-identified archive of colposcopic images obtained from a prospective tertiary-level colposcopy clinic. Institutional ethics committee approval was granted (reference number withheld pending peer review). Informed consent was obtained from all patients whose images were included in the archive, in accordance with institutional policies. The study was conducted in compliance with the Declaration of Helsinki.

Consent for Publication

Not applicable. This manuscript does not contain data from any individual person in any identifiable form. All images are fully de-identified.

Availability of Data and Materials

The colposcopic image dataset used in this study was obtained from a clinical archive and contains sensitive patient data; it is therefore not publicly available. Requests for access to the de-identified dataset may be directed to the corresponding author and are subject to institutional data governance approval. The source code for HDFF-Net, feature extraction, and stacked ensemble will be made available on the corresponding author's institutional repository upon acceptance of this manuscript.

Competing Interests

The authors declare that they have no competing interests.

Funding

This work received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Authors' Contributions

AA conceptualised the study, designed and implemented the HDFF-Net architecture, conducted all experiments, and drafted the manuscript. BB contributed to the clinical dataset curation, provided expert colposcopic labels, and critically revised the manuscript for clinical accuracy. CC contributed to the statistical analysis framework and revised the manuscript. All authors read and approved the final manuscript.

Acknowledgements

The authors thank the colposcopy clinic staff for their cooperation in dataset curation and the two expert colposcopists who provided consensus TZ-type labels for the dataset.

REFERENCES

[1] Sung H, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2021;71(3):209–249.

[2] Arbyn M, et al. Estimates of incidence and mortality of cervical cancer in 2018: a worldwide analysis. Lancet Glob Health. 2020;8:e191–e203.

[3] World Health Organization. Global strategy to accelerate the elimination of cervical cancer as a public health problem. Geneva: WHO; 2020.

[4] Cox JT. Epidemiology of cervical intraepithelial neoplasia. Clin Obstet Gynaecol. 1995;9(1):1–37.

[5] Bornstein J, et al. 2011 colposcopic terminology of the IFCPC. Obstet Gynecol. 2012;120(1):166–172.

[6] Kyrgiou M, et al. Obstetric outcomes after conservative treatment for cervical intraepithelial lesions. Cochrane Database Syst Rev. 2017;9:CD012847.

[7] Massad C, et al. Interobserver agreement in colposcopy. Gynecol Oncol. 2006;100:59–65.

[8] Ferris TM, et al. Agreement among colposcopists for the assessment of transformation zone type. J Low Genit Tract Dis. 2016;20:321–325.

[9] Duda RO, Hart PE, Stork DG. Pattern Classification. 2nd ed. New York: Wiley; 2001.

[10] Haralick RM, Shanmugam K, Dinstein I. Textural features for image classification. IEEE Trans Syst Man Cybern. 1973;SMC-3:610–621.

[11] Ojala T, Pietikainen M, Maenpaa T. Multiresolution gray-scale and rotation invariant texture classification with LBP. IEEE Trans Pattern Anal Mach Intell. 2002;24:971–987.

[12] Dalal N, Triggs B. Histograms of oriented gradients for human detection. Proc CVPR. 2005:886–893.

[13] Manjunath BS, Ma WY. Texture features for browsing and retrieval of image data. IEEE Trans Pattern Anal Mach Intell. 1996;18:837–842.

[14] Simmons RM, Gierach B, Colegrove A. SVM classification of colposcopic image features. Comput Biol Med. 2016;72:93–101.

[15] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. Proc ICLR. 2015.

[16] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Proc CVPR. 2016:770–778.

[17] Szegedy C, et al. Rethinking the inception architecture for computer vision. Proc CVPR. 2016:2818–2826.

[18] Miyagi Y, Takehara Y, Nagayama T. Application of deep learning to colposcopic cervical image classification. J Gynecol Oncol. 2020;31:e96.

[19] Wilhelm T, Kurzeder R, Fehm MG. Patch-based deep learning for colposcopy adequacy. Proc MICCAI. 2022:211–220.

[20] [Authors]. Type-based colposcopic image classification using hybrid handcrafted descriptors and deep transfer learning. [Prior paper, same authors]; 2024.

[21] Hu J, Shen L, Sun G. Squeeze-and-excitation networks. Proc CVPR. 2018:7132–7141.

[22] Tan M, Le Q. EfficientNet: Rethinking model scaling for convolutional neural networks. Proc ICML. 2019:6105–6114.

[23] Woo S, Park J, Lee J-Y, Kweon IS. CBAM: Convolutional block attention module. Proc ECCV. 2018:3–19.

[24] Howard J, Ruder S. Universal language model fine-tuning for text classification. Proc ACL. 2018:328–339.

[25] Lindeberg T. Scale-Space Theory in Computer Vision. Dordrecht: Kluwer Academic; 1994.

[26] Huang X, Zhao G. Multi-scale local binary pattern for face recognition. Proc FG. 2013:1–6.

[27] Guo Z, Zhang L, Zhang D. A completed modeling of local binary pattern operator. IEEE Trans Image Process. 2010;19:1657–1663.

[28] Zhu Q, Avidan S, Yeh M, Cheng K. Fast human detection using a cascade of histograms of oriented gradients. Proc CVPR. 2006:1491–1498.

[29] Cheng MJ, Wang J, Feng J. Multi-scale HOG for scene classification. Proc ICIP. 2013:3801–3805.

[30] Soh L-K, Tsatsoulis C. Texture analysis of SAR sea ice imagery using GLCM. IEEE Trans Geosci Remote Sens. 1999;37:780–795.

[31] Schlemper J, et al. Attention gating for improving U-Net for retinal vessel segmentation. Proc MIDL. 2019:1–10.

[32] Guo Z, et al. Channel attention and squeeze-and-excitation for histopathology grading. Proc MICCAI. 2020:350–359.

[33] Dosovitskiy A, et al. An image is worth 16×16 words: Transformers for image recognition at scale. Proc ICLR. 2021.

[34] Ramachandram D, Taylor GW. Deep multimodal learning: A survey. IEEE Signal Process Mag. 2017;34:96–108.

[35] Radford A, et al. Learning transferable visual models from natural language supervision. Proc ICML. 2021:8748–8763.

[36] Zhang Z, et al. Biomedical and clinical English model packages for the Stanza Python NLP library. J Am Med Inform Assoc. 2021;28:1892–1899.

[37] Tan M, Pang R, Le QV. EfficientDet: Scalable and efficient object detection. Proc CVPR. 2020:10781–10790.

[38] Raghu M, Zhang C, Kleinberg J, Bengio S. Transfusion: Understanding transfer learning for medical imaging. Adv NeurIPS. 2019:3347–3357.

[39] Wolpert DH. Stacked generalization. Neural Netw. 1992;5(2):241–259.

[40] Chen T, Guestrin C. XGBoost: A scalable tree boosting system. Proc KDD. 2016:785–794.

[41] Esteva A, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017;542:115–118.

[42] He H, Garcia EA. Learning from imbalanced data. IEEE Trans Knowl Data Eng. 2009;21:1263–1284.

[43] Hendrycks D, Gimpel K. Gaussian error linear units (GELUs). arXiv:1606.08415; 2016.

[44] Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture. Proc CVPR. 2016.

[45] Batista G, Prati RC, Monard MC. A study of the behaviour of several methods for balancing machine learning training data. ACM SIGKDD Explor. 2004;6:20–29.

[46] Liu Z, et al. Swin Transformer: Hierarchical vision transformer using shifted windows. Proc ICCV. 2021:10012–10022.

[47] McMahan B, et al. Communication-efficient learning of deep networks from decentralized data. Proc AISTATS. 2017:1273–1282.

[48] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic minority over-sampling technique. J Artif Intell Res. 2002;16:321–357.

[49] Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. Adv NeurIPS. 2017:4765–4774.

[50] Selvaraju RR, et al. Grad-CAM: Visual explanations from deep networks via gradient-based localisation. Proc ICCV. 2017:618–626.

[51] Powers DMW. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. J Mach Learn Technol. 2011;2:37–63.

[52] DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves. Biometrics. 1988;44:837–845.

[53] Kingma DP, Ba J. Adam: A method for stochastic optimization. Proc ICLR. 2015.

[54] Loshchilov I, Hutter F. Decoupled weight decay regularization. Proc ICLR. 2019.

[55] Litjens G, et al. A survey on deep learning in medical image analysis. Med Image Anal. 2017;42:60–88.

[56] Pedregosa F, et al. Scikit-learn: Machine learning in Python. J Mach Learn Res. 2011;12:2825–2830.

[57] Lemaitre G, Nogueira F, Aridas CK. Imbalanced-learn: A Python toolbox to tackle the curse of imbalanced datasets in machine learning. J Mach Learn Res. 2017;18:1–5.

[58] Vaswani A, et al. Attention is all you need. Adv NeurIPS. 2017:5998–6008.

[59] Abadi M, et al. TensorFlow: A system for large-scale machine learning. Proc USENIX OSDI. 2016.

[60] Rajpurkar P, Lungren M, Ng A. CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning. arXiv:1711.05225; 2017.

[61] Prandzhev GD, Gortchev GA, Dimitrov DD, Miltchev RI, Rangelov D, Tomov ST. Artificial intelligence analysis of the transformation zone of the uterine cervix: colposcopic cervical categorization and integration of deep learning. J Biomed Clin Res. 2025;18:47–53. https://doi.org/10.3897/jbcr.e144006

[62] Cao Y, Ma H, Fan Y, Liu Y, Zhang H, Cao C, Yu H. A deep learning-based method for cervical transformation zone classification in colposcopy images. Technol Health Care. 2023;31(2):527–538. https://doi.org/10.3233/THC-220141

[63] Behera SK, Mahakud R, Panigrahi M, Sethy PK, Pati R. Cervical transformation zone segmentation and classification based on improved Inception-ResNet-V2 using colposcopy images. Int Ophthalmol. 2024;44(1):174. https://doi.org/10.1007/s10792-024-03058-0

[64] Hong Z, Xiong J. Lightweight low-rank adaptation Vision Transformer framework for cervical cancer detection and cervix type classification. Bioengineering (Basel). 2024;11(5):468. https://doi.org/10.3390/bioengineering11050468

[65] Al-Hejri AM, Al-Tam RM, Sable AH, et al. A hybrid vision transformer with ensemble CNN framework for cervical cancer diagnosis. BMC Med Inform Decis Mak. 2025;25(1):411. https://doi.org/10.1186/s12911-025-03250-x

[66] Hemajothi S, Ravi VR, Girinandanaa G, Joycerobega PD, Moncia K. Transformative advances in cervical cancer diagnosis leveraging colposcopy imaging with ViT. Proc 2024 Int Conf Recent Adv Electr Electron Ubiquitous Commun Comput Intell (RAEEUCCI). IEEE; 2024:1–6.

[67] Nirmala G, Nayudu PP, Kumar AR, Sagar R. Automatic cervical cancer classification using adaptive vision transformer encoder with CNN for medical application. Pattern Recognit. 2024;156:111201. https://doi.org/10.1016/j.patcog.2024.111201

[68] Dakshayani Himabindu D, Laxmi Lydia E, Rajesh MV, et al. Leveraging Swin Transformer with ensemble of deep learning model for cervical cancer screening using colposcopy images. Sci Rep. 2025;15:23841. https://doi.org/10.1038/s41598-025-90415-3

[69] Chen X, Pu X, Chen Z, Li L, Zhao KN, Liu H, Zhu H. Application of EfficientNet-B0 and GRU-based deep learning on classifying the colposcopy diagnosis of precancerous cervical lesions. Cancer Med. 2023;12:14706–14717. https://doi.org/10.1002/cam4.6080

[70] Mukku L, Thomas J. CMT-CNN: colposcopic multimodal temporal hybrid deep learning model to detect cervical intraepithelial neoplasia. Int J Adv Intell Inform. 2024;10(2):183–198. https://doi.org/10.26555/ijain.v10i2

[71] Dash M, et al. Deep learning descriptor hybridization with feature reduction for accurate cervical cancer colposcopy image classification. arXiv:2405.01600; 2024. https://arxiv.org/abs/2405.01600

[72] Karaboga D, et al. Multi-modal graph neural networks for colposcopy data classification and visualization. Diagnostics (Basel). 2025;15(9):1102. https://doi.org/10.3390/diagnostics15091102

[73] Mathivanan SK, Francis D, Srinivasan S, Khatavkar V, Shah MA. Enhancing cervical cancer detection and robust classification through a fusion of deep learning models. Sci Rep. 2024;14:10812. https://doi.org/10.1038/s41598-024-61402-3

[74] Li B, Chen L, Sun C, et al. Leveraging deep learning for early detection of cervical cancer and dysplasia in China using UNet++ and RepVGG networks. Front Oncol. 2025;15:1624111. https://doi.org/10.3389/fonc.2025.1624111

[75] Xing M, et al. Achieving flexible fairness metrics in federated medical imaging. Nat Commun. 2025;16:3219. https://doi.org/10.1038/s41467-025-58549-0