DiffAFCA-Net: A Diffusion-Based Framework for Cervical Precancerous Cell Classification with Adaptive Fine-Grained Channel Attention

1st Tianyu Shi
School of Information Science and Engineering
Shenyang Ligong University
Shenyang City, P.R. China
ty.shi@sylu.edu.cn

2nd Xu Ma
School of Information Science and Engineering
Shenyang Ligong University
Shenyang City, P.R. China
maxu20011005@163.com

3rd Zhimin Liu
Department of Pathology
The Second Affiliated Hospital of Dalian Medical University
Dalian City, P.R. China
liuzhimin@dmu.edu.cn

4th Xiaoyun Liu
School of Science
Shenyang Ligong University
Shenyang City, P.R. China
liuxy@imr.ac.cn

5th Qingkun Guo*
School of Information Science and Engineering
Shenyang Ligong University
Shenyang City, P.R. China
bettergkun@hotmail.com

6th Meng Wang*
School of Electronic and Information Engineering
Liaoning Technical University
Xingcheng City, P.R. China
wangmeng_dx@lntu.edu.cn

***Abstract*—Cervical cancer remains a significant global health challenge, with early detection critically dependent on accurate classification of precancerous lesions in cytological images. However, conventional deep learning approaches often struggle with robustness issues under complex noise interference, such as staining variations and cellular overlaps, and exhibit limited sensitivity to fine-grained morphological features essential for distinguishing subtle pathological changes. To address these gaps, this paper proposes DiffAFCA-Net, a novel diffusion-based framework that integrates adaptive fine-grained channel attention for enhanced feature representation in cervical cell classification. The model leverages a denoising diffusion probabilistic process to implicitly learn robust latent representations, while an adaptive attention mechanism dynamically captures both global and local discriminative features at multiple granularities, enabling precise emphasis on critical cellular structures like nuclear-cytoplasmic ratios and chromatin distribution. Experiments on multiple cervical cell datasets demonstrate that DiffAFCA-Net achieves superior performance compared to state-of-the-art methods, with notable improvements in classification accuracy, F1-score, and robustness under noisy conditions. The results highlight the potential of combining diffusion models with fine-grained attention for reliable automated screening, offering a scalable solution for clinical applications where label noise and domain shifts are prevalent.**

Keywords—Diffusion Probabilistic Model, Medical Image Classification, Cervical Precancerous Cell, Attention Mechanisms

I. INTRODUCTION

Cervical cancer continues to be a major public health concern, accounting for a significant proportion of cancer-related morbidity and mortality among women worldwide, with pronounced disparities in healthcare access and screening coverage across different regions [1]. Early detection through cytological examination, has proven effective in reducing incidence and mortality rates in high-income countries [2]. However, traditional screening methods face inherent limitations, including subjective interpretation, inter-observer variability, and high false-negative rates, which are exacerbated in resource-constrained settings where expertise and infrastructure are limited [3]. The advent of whole-slide imaging and computer-aided diagnosis systems has spurred the development of deep learning-based approaches to automate cell classification, yet key challenges persist in terms of robustness, generalization, and fine-grained feature representation [4].

Convolutional neural networks (CNNs), such as ResNet [5], have demonstrated considerable success in cervical cell image analysis by leveraging hierarchical feature learning. For instance, Zhang *et al.* [6] achieved notable accuracy using CNNs on the Herlev dataset, but these models often struggle with complex background noise, staining variations, and cell overlapping, leading to performance degradation in real-world scenarios. To enhance feature discrimination, attention mechanisms have been integrated into deep networks[7]. Yet, existing methods primarily employ static or global attention strategies [8], which may insufficiently capture subtle morphological details, such as nuclear-cytoplasmic (N/C) ratio anomalies or chromatin distribution patterns, critical for distinguishing low-grade and high-grade squamous intraepithelial lesions (LSIL/HSIL) [9].

Recent years have witnessed the rise of transformer-based architectures[10][11], which excel at modeling long-range dependencies through self-attention. These models have shown promising results in medical image classification tasks by leveraging large-scale receptive fields and multi-scale feature aggregation [12]. Nevertheless, transformers often require extensive computational resources and large datasets for training, limiting their applicability in data-scarce medical domains. Additionally, generative models, particularly denoising diffusion probabilistic models (DDPMs) [13], have emerged as a powerful paradigm for learning robust data distributions. DDPMs progressively corrupt input data with Gaussian noise in the forward process and learn to reverse this noise through a Markov chain, enabling high-fidelity image generation and enhanced feature representation [14]. In medical imaging, diffusion models have been applied to tasks like denoising [15] and classification [16], offering improved robustness to noise and domain shifts. However, current diffusion-based approaches often overlook the integration of fine-grained, adaptive attention mechanisms, resulting in

This work was supported by the National Natural Science Foundation of China (Grant No. 62506243, 82572327), the Shenyang Ligong University High-level Talent Cultivation Research Support Program (Grant No. 1010147001232), the Liaoning Provincial Science and Technology Planning Project (Grant No. 2025080007-JH3/107), and the Education Department of Liaoning Provincial Youth Scientist Fund (Grant No. LJ212510144013).

suboptimal utilization of pathological cues for precise cell classification.

To address these gaps, we propose DiffAFCA-Net, a novel framework that synergizes diffusion models with an adaptive fine-grained channel attention mechanism for cervical precancerous cell classification. Our key contributions are threefold: (1) We introduce an adaptive fine-grained attention module that dynamically fuses global and local channel features using cross-correlation operations, enhancing sensitivity to critical pathological characteristics. (2) We leverage a diffusion-based pipeline to implicitly model robust feature representations, mitigating the impact of complex noise and improving generalization across diverse datasets. (3) Extensive experiments on multiple cervical cell image datasets demonstrate that DiffAFCA-Net outperforms state-of-the-art methods in accuracy, F1-score, and robustness, while maintaining computational efficiency. By combining the strengths of diffusion models and fine-grained attention, our approach provides a scalable solution for automated screening, with potential applications in clinical workflows where reliability and interpretability are paramount.

II. METHODOLOGY

The framework of DiffAFCA-Net is illustrated in Figure 1. Given an input cervical cell image x, shallow features are first extracted using a prior network ρ_{cls}. These features are then passed to the Adaptive Fine-Grained Channel Attention (AFCA) module, which generates a global feature map G^* and a local feature map L^*. Both maps are fed into a feature fusion module to produce global prior $\hat{y}_g$, local prior $\hat{y}_l$, and a dual prior f that combines global and local representations. During training, the diffusion model is applied to the clean sample y_0 under different prior conditions, producing noisy variables y_t^g, y_t^l, and y_t corresponding to global, local, and dual priors, respectively. These noisy representations are projected into a latent space and integrated with image features from the encoder ρ_{prior} within a denoising U-Net. The model is optimized using a multi-task loss that balances noise prediction accuracy and feature consistency.

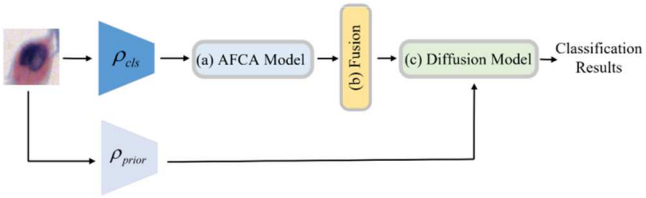

Fig. 1. End-to-end architecture of DiffAFCA-Net framework, illustrating the diffusion-driven pipeline with adaptive fine-grained attention.

A. Adaptive Fine-Grained Channel Attention Module

The AFCA module is designed to capture both global and local channel-wise dependencies adaptively, enabling the model to emphasize discriminative features such as nuclear-cytoplasmic ratio and chromatin distribution. As shown in Figure 2, the module leverages cross-correlation operations to model interactions between global and local features at different granularities. Given an input feature map $F \in R^{C\times H\times W}$, where C, H, and W denote channels, height, and width, respectively, the process begins with global average pooling (GAP) to obtain a channel descriptor U:

$$U_n = GAP(F_n) = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} F_n(i,j) \quad (1)$$

where $F_n(i,j)$ is the pixel value at position (i, j) in the n-th channel. To model local channel interactions, a band matrix $B = [b_1, b_2, b_3, \cdots, b_k]$ is applied using 1D convolution, producing local features $U_{lc} = \sum_{l=1}^{k} U \cdot b_l$. For global dependencies, a diagonal matrix $D = \left[d_1, d_2, d_3, \cdots, d_c\right]$ is used via 2D convolution, yielding global features $U_{gc} = \sum_{i=1}^{c} U \cdot d_i$. The correlation between global and local features is computed as:

$$M = U_{gc} \cdot U_{lc}^{T} \quad (2)$$

where M is the correlation matrix. Adaptive fusion is achieved through learnable scaling factors, generating final weights W as:

$$W = \sigma(\sigma(\theta)\times\sigma(U_{gc}^{a}) + (1-\sigma(\theta))\times\sigma(U_{ic}^{a})) \quad (3)$$

with σ denoting the sigmoid function. The output feature map F^* is obtained via element-wise multiplication:

$$F^* = W \otimes F \quad (4)$$

This design enables precise emphasis on pathological cues such as nuclear boundaries and chromatin patterns.

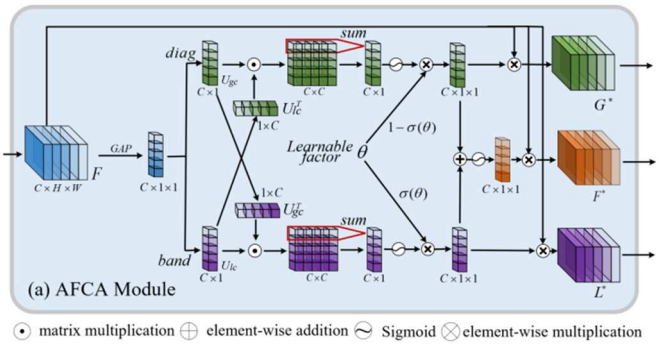

Fig. 2. Structure of AFCA module, demonstrating the cross-correlation based fusion of global and local channel features. The module employs band and diagonal matrices for multi-granularity feature interaction, with adaptive weighting through learnable parameters.

B. Feature Fusion Module

To handle scale variations and semantic interactions in cervical cell images, a feature fusion module integrates global features G^* local features L^*, and semantic features F^* from the AFCA output. For G^*, channel attention weights are derived using global average and max pooling, followed by a shared multilayer perceptron (MLP):

$$\hat{G} = (\sigma(\mathrm{MLP}(AvgPool(G^*) + \mathrm{MLP}(MaxPool(G^*))))) \otimes G^* \quad (5)$$

where $\otimes$ denotes element-wise multiplication. Similarly, for L^*, spatial attention is applied via concatenation of pool outputs and a 7×7 convolution:

$$\hat{L} = (\sigma(f^{7\times7}(Concat[AvgPool(L^*), MaxPool(L^*)]))) \otimes L^* \quad (6)$$

Semantic features F^* are reduced using a 1×1 convolution and global average pooling: $\tilde{F} = Avgpool(f^{1\times1}(F^*))$. The fused feature $\hat{F}$ is obtained by concatenating and processing through a 1×1convolution:

$$\hat{F} = f^{1\times1}(Concat[G^*, L^*, \tilde{F}]) \quad (7)$$

Two parallel branches with depth-wise 3×3 convolutions and Layer Normalization (LN) further refine the features,

producing the final output $F^{'}$. This module enhances the model's ability to represent complex morphological variations.

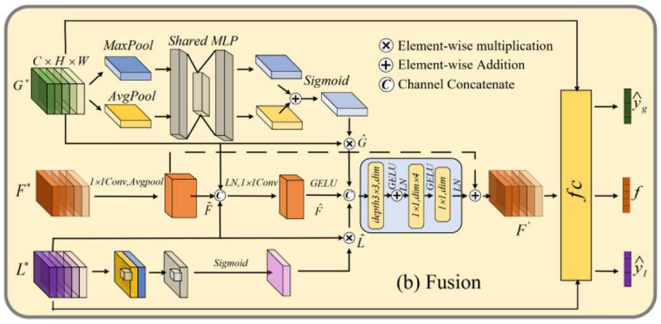

Fig. 3. Feature fusion module with multi-branch aggregation, showing the integration of global, local, and semantic features through parallel processing paths. The design includes channel-spatial attention mechanisms and residual connections for enhanced representation learning.

C. Diffusion Model Integration

The diffusion model follows the DDPM framework [13], which consists of a forward noise-adding process and a reverse denoising process. In the forward process, the clean label representation y_0 is progressively corrupted with Gaussian noise over T steps , producing a sequence of noisy variables $\{y_1, \cdots, y_T\}$:

$$q(y_t|y_0) = N(y, \sqrt{\overline{\alpha}_t} y_0, (1-\overline{\alpha}_t)I) \tag{8}$$

where, $\overline{\alpha}_t = \prod_{i-1}^{t} \alpha_i$, and $\alpha_i = 1 - \beta_i$ with β_i being the noise variance added to the image at time step i. The noisy sample y_t is obtained using reparameterization:

$$y_t = \sqrt{\overline{\alpha}_t} y_0 + \sqrt{(1-\alpha_t)}\varepsilon, \quad \varepsilon \sim N(0, I) \tag{9}$$

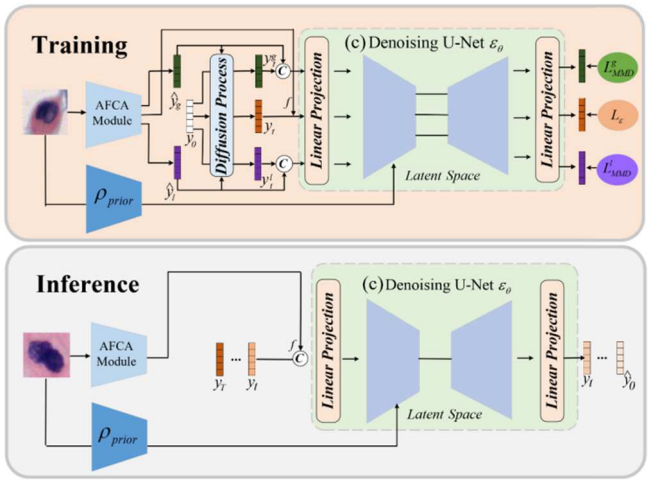

Fig. 4. Diffusion model integration workflow, depicting the Markov chain process of forward noise addition and reverse denoising conditioned on AFCA-generated priors.

In the reverse process, a U-Net denoising network ε_θ predicts the noise component conditioned on the priors from the AFCA module. The reverse step is given by:

$$y_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(y_t - \frac{1-\alpha_t}{\sqrt{1-\overline{\alpha}_t}}\varepsilon_\theta\left(y_t, t\right)\right) + \delta_t z, \; z \sim N(0, I) \tag{10}$$

where δ_t is the variance schedule. This conditioning mechanism enhances the model's focus on diagnostically relevant regions.

D. Training Loss and Optimization

A multi-task loss function is employed to optimize the model, combining constraints for global features, local features, and classification output. Let ε_g, ε_l and ε_f denote the noise estimates for the global, local, and dual prior branches, respectively. The Maximum Mean Discrepancy (MMD) loss [14] is used to regularize the global and local branches, measuring the similarity between the sampled noise $\epsilon$and the estimated noise:

$$L_g(\epsilon \| \epsilon_g) = K(\epsilon, \epsilon) - 2K(\epsilon_g, \epsilon) + K(\epsilon_g, \epsilon_g), \tag{11}$$

$$L_l(\epsilon \| \epsilon_l) = K(\epsilon, \epsilon) - 2K(\epsilon_l, \epsilon) + K(\epsilon_l, \epsilon_l), \tag{12}$$

where $K(\cdot,\cdot)$ denotes a positive-definite kernel function in a reproducing kernel Hilbert space (RKHS). For the classification task, the mean squared error (MSE) loss is applied to the dual prior branch:

$$L_\varepsilon = \|\varepsilon - \varepsilon_f\|^2 \tag{13}$$

The total loss is a weighted sum:

$$L_{total} = \omega(L_g + L_l) + L_\varepsilon \tag{14}$$

where ω is a hyperparameter that balances the components. This loss design ensures robust feature learning and effective noise prediction.

E. Inference Stage

During inference, given an input image x, the AFCA module generates the priors $\hat{y}_g$ and $\hat{y}_l$. The diffusion model starts from a random noise vector y_T and iteratively denoises it using the U-Net conditioned on the priors and image features, producing the final prediction $\hat{y}_0$. This process enhances stability and accuracy in real-world scenarios.

III. EXPERIMENTAL EVALUATION

In this section, we present a comprehensive evaluation of the proposed DiffAFCA-Net framework on multiple cervical cell image datasets. We first describe the datasets used, followed by the experimental setup, including implementation details and evaluation metrics. Next, we compare our method with state-of-the-art approaches quantitatively and qualitatively, and conduct ablation studies to validate the contribution of each component. All experiments are designed to assess the model's classification accuracy, robustness, and generalization capability under realistic clinical scenarios.

A. Datasets

We used three publicly available datasets and one private clinical dataset for evaluation, ensuring diversity in image sources, staining protocols, and class distributions.

SIPaKMeD [17] is a publicly available dataset containing 2,029 single-cell images extracted from 966 cervical smears. The images are categorized into five distinct classes: Superficial-Intermediate, Parabasal, Koilocytotic, Dyskeratotic, and Metaplastic cells. This dataset presents challenges such as staining variations and subtle

morphological differences between different cell types, making it suitable for evaluating fine-grained classification capabilities.

DSCC [18] comprises 10,294 images with 128×128 pixel resolution, distributed across five categories: Normal (3,350 images), ASC-US (1,744 images), ASC-H (1,600 images), LSIL (2,150 images), and HSIL (1,450 images). The images are derived from ThinPrep liquid-based cytology samples stained with H&E, featuring complex backgrounds and cell overlaps. The significant class imbalance and preparation artifacts in this dataset test model robustness under realistic conditions.

CerviClin is a private clinical dataset consisting of 7,671 images across four categories: ASC-H (1,915 images), ASCUS (1,931 images), HSIL (1,926 images), and LSIL (1,899 images). The images were acquired using clinical-grade digital slide scanners under routine diagnostic settings. Each image has a resolution of 128×128 pixels. All images were annotated by board-certified cytopathologists following the Bethesda classification guidelines, with each label confirmed through multi-reader review to ensure diagnostic consistency. The dataset's design captures real-world heterogeneity in staining intensity, focus, and illumination, making it a rigorous benchmark for evaluating the clinical applicability and domain-adaptation capacity of automated cervical-cell analysis algorithms.

B. Experimental Setup

Experiments were conducted on a system with an NVIDIA RTX 4080S GPU, using PyTorch 1.10.0 and CUDA 12.8. The model was trained with a batch size of 32, Adam optimizer [19] (learning rate = 2e-4, $\beta_1 = 0.9$), and a linear noise schedule for the diffusion process (β_1 = 1e-4 to $\beta_T = 0.02$ over T = 1000 steps). The diffusion model was integrated with a ResNet-18 [5] backbone for feature extraction. Training included a 15-epoch warm-up phase for the AFCA module, followed by joint optimization for 100 epochs. Five-fold cross-validation was applied to all datasets, and results are reported as mean ± standard deviation over three runs [20].

C. Comparative Results

Results on SIPaKMeD Dataset: As demonstrated in Table I, the proposed DiffAFCA-Net achieves state-of-the-art performance across all evaluation metrics. Our method attains an accuracy of 97.95% and an F1-score of 98.02%, outperforming the strongest baseline, HiFuse, by a margin of 1.71% in accuracy and 1.78% in F1-score. The Cohen's Kappa coefficient of 97.52% indicates superior inter-class consistency, while the high true positive rate and true negative rate underscore the model's balanced sensitivity and specificity. Notably, DiffAFCA-Net surpasses diffusion-based baseline DiffMIC by 2.60% in accuracy, highlighting the efficacy of integrating adaptive fine-grained attention with diffusion processes. The AUC value of 99.02% further confirms the model's robust discriminative capability, with ROC curves concentrated near the top-left corner for all five classes. These results validate the model's ability to capture fine-grained morphological features, such as nuclear boundaries and chromatin patterns, under staining variations and subtle cellular differences.

Results on DSCC Dataset: As summarized in Table II, DiffAFCA-Net maintains competitive performance on the DSCC dataset, achieving an accuracy of 93.70% and an F1-score of 93.29%. This represents a significant improvement over transformer-based methods, such as Swin Transformer and ViT, with accuracy gains of 21.60% and 20.25%, respectively. The model also outperforms recent approaches like DiffMIC and HiFuse, demonstrating better generalization to multi-class scenarios with imbalanced distributions. The Kappa coefficient of 90.49% reflects high agreement with expert annotations, while the TNR and AUC indicate strong robustness to background clutter and noise. The diffusion prior effectively mitigates artifacts from liquid-based cytology preparations, while the AFCA module enhances focus on critical regions, such as irregular nuclei in SIL categories.

TABLE I. PERFORMANCE COMPARISON OF THE SIPAKMED DATASET.

Method	Acc	F1	Kappa	TPR	TNR	AUC
ResNet[5]	89.84	89.88	87.29	89.81	97.46	98.38
ViT[10]	76.49	76.41	70.62	76.65	76.77	94.36
Swin[11]	95.92	95.90	94.89	95.92	95.92	**99.80**
ConvNeXt[21]	92.35	92.28	89.05	93.90	98.36	99.23
DiffMIC[14]	95.35	96.54	96.22	97.69	99.40	99.33
MedMamba[22]	94.74	94.72	93.42	94.73	98.68	99.31
HiFuse[23]	96.24	96.24	95.29	96.23	99.06	99.60
Ours	**97.95**	**98.02**	**97.52**	**98.01**	**99.52**	99.02

TABLE II. PERFORMANCE COMPARISON OF THE DSCC DATASET.

Method	Acc	F1	Kappa	TPR	TNR	AUC
ResNet[5]	80.32	77.31	69.49	76.94	89.99	91.47
ViT[10]	73.45	68.43	58.62	69.12	69.2	87.83
Swin[11]	72.1	71.77	58.66	72.1	72.1	92.44
ConvNeXt[21]	79.46	71.59	67.46	73,26	89.55	92.43
DiffMIC[14]	93.57	92.98	83.01	92.95	96.78	96.3
MedMamba[22]	82.34	81.31	73.37	82.43	91.66	94.34
HiFuse[23]	82.86	81.38	73.95	81.99	91.75	94.54
Ours	**93.70**	**93.29**	**90.49**	**93.34**	**96.90**	**97.20**

Results on CerviClin Dataset: On the private clinical dataset (Table III), DiffAFCA-Net achieves the highest accuracy and F1-score, confirming its practicality in real-world settings. The model surpasses ResNet-50 by 9.83% and DiffMIC by a narrow margin, while maintaining superior Kappa and TPR values. The TNR of 91.11% and AUC of 90.71% highlight its ability to reduce false positives in clinical workflows, where specificity is critical for triage. These results underscore the model's adaptability to domain shifts, such as equipment variations and staining inconsistencies, leveraging the diffusion framework's noise robustness and the attention mechanism's feature refinement. The consistent performance across datasets aligns with clinical needs for reliable screening tools, particularly in distinguishing challenging cases like ASC-H vs. HSIL.

TABLE III. PERFORMANCE COMPARISON OF THE CERVICLIN DATASET.

Method	Acc	F1	Kappa	TPR	TNR	AUC
ResNet[5]	63.68	63.73	51.54	64.01	87.85	87.14
ViT[10]	57.73	57.28	43.61	57.65	57.46	85.77
Swin[11]	67.84	65.48	45.60	67.84	67.84	**92.81**
ConvNeXt[21]	60.46	60.04	47.28	60.49	86.82	89.69
DiffMIC[14]	73.56	73.58	59.21	73.56	91.19	86.00
MedMamba[22]	69.74	57.62	59.59	69.99	89.85	89.52
HiFuse[23]	68.88	69.61	58.41	69.12	89.54	91.11
Ours	**73.51**	**73.48**	**64.12**	**73.53**	**91.11**	90.71

D. Ablation Studies

To dissect the contribution of each component, we conducted ablation experiments on the SIPaKMeD dataset, as summarized in Table IV. We evaluated four variants:

Basic: ResNet-18 without diffusion or attention modules.

C1: Basic + diffusion process.

C2: C1 + AFCA module.

C3: C2 + feature fusion module.

The results show progressive improvements: adding diffusion (C1) increased accuracy by 2.16% over Basic, highlighting its noise robustness. Incorporating AFCA (C2) boosted F1-score by 1.61%, validating its role in fine-grained feature capture. The fusion module (C3) further enhanced accuracy by 2.88%, demonstrating effective multi-scale integration. The full model achieved the best performance, confirming the synergy between components.

TABLE IV. ABLATION STUDY ON SIPAKMED DATASET

Model	Diff	FCA	Fusion	Accuracy	F1-score
Basic	-	-	-	92.60	92.59
C1	√	-	-	94.76	95.65
C2	√	√	-	95.06	97.26
C3	√	√	√	97.94	98.04

E. Visualization and Analysis

The ROC curves for the DSCC dataset, as illustrated in Figure 5, demonstrate that DiffAFCA-Net achieves superior discriminative performance. In detail, the ROC analysis reveals that DiffAFCA-Net maintains consistent performance for each class. For the Normal class, the curve shows high sensitivity and specificity, reducing the risk of false negatives. For ASC and SIL classes, which are often challenging to differentiate due to subtle morphological differences, the model exhibits steep curves, indicating rapid improvement in true positive rates with minimal increase in false positives. The superiority of DiffAFCA-Net can be attributed to its integrated components, which enhance feature representation and long-range dependency capture.

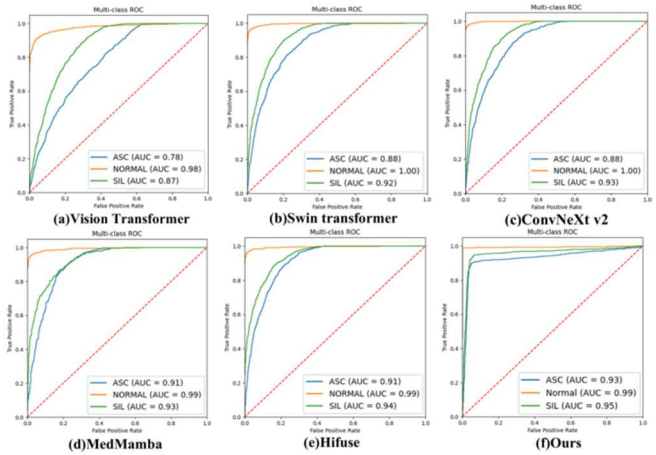

Fig. 5. ROC curves on DSCC dataset, comparing the classification performance of DiffAFCA-Net against state-of-the-art methods across five cervical cell categories.

IV. DISCUSSION AND ANALYSIS

A. Performance Interpretation and Clinical Relevance

The experimental results demonstrate that DiffAFCA-Net achieves state-of-the-art performance across multiple datasets, with particular strengths in handling real-world challenges inherent to cervical cytology images. The superior performance can be attributed to several key factors rooted in the model's architectural design and learning mechanism.

The integration of diffusion models provides a robust framework for learning stable feature representations under noisy conditions. Unlike conventional discriminative models that may overfit to specific artifact patterns in training data, the denoising process in DiffAFCA-Net enables the model to develop inherent resistance to various types of noise commonly encountered in clinical settings, including staining variations, cellular overlaps, and background debris. This is particularly evident in the model's strong performance on the DSCC dataset, which contains substantial background complexity and preparation artifacts.

The AFCA mechanism plays a crucial role in capturing subtle morphological features that are clinically significant for cervical cell classification. Pathologists typically rely on fine-grained characteristics such as nuclear-cytoplasmic ratio abnormalities, chromatin distribution patterns, and nuclear membrane irregularities to differentiate between normal and abnormal cells, particularly when distinguishing between low-grade and high-grade lesions. The AFCA module's ability to dynamically weight channel features based on both local and global contexts allows the model to focus on these discriminative patterns, mimicking the expert pathologist's visual search strategy.

The feature fusion module further enhances the model's capability to integrate multi-scale information, which is essential for handling the substantial size variations observed in cervical cells across different pathological states. This multi-scale approach enables the network to simultaneously capture both cell-level contextual information and subcellular details, providing a comprehensive representation that aligns with the hierarchical examination process used in manual cytological assessment.

B. Computational Efficiency and Practical Deployment Considerations

While DiffAFCA-Net demonstrates superior classification performance, its practical deployment requires consideration of computational requirements. The diffusion process introduces additional computational overhead compared to standard classification networks. During training, the model requires 18-24 hours on a single RTX 4080 Super GPU to converge on the largest dataset, while inference for a batch of 32 images takes approximately 0.8 seconds. This represents a 3-5× increase in inference time compared to ResNet-50 baselines but remains within practical limits for clinical screening applications where throughput requirements are typically measured in minutes per case rather than seconds.

Several strategies can optimize computational efficiency for real-world deployment. Knowledge distillation techniques could transfer the learned representations to a lighter-weight student network, reducing inference time while preserving most of the performance gains. Additionally, progressive distillation methods specific to diffusion models could compress the multi-step denoising process into fewer steps, significantly accelerating inference without substantial accuracy loss.

The model's memory footprint of 6.2 GB during training and 2.1 GB during inference makes it suitable for deployment on standard clinical workstations. For resource-constrained environments, quantization techniques could further reduce memory requirements, though this may necessitate careful calibration to maintain the precision required for medical diagnostics.

C. Limitations and Failure Case Analysis

Despite the strong overall performance, analysis of misclassified cases reveals several limitations that warrant attention in future work. The most common errors occur in borderline cases where cellular abnormalities are subtle or ambiguous, particularly in distinguishing between ASC-H and early HSIL categories. These challenging cases often also pose difficulties for human experts, suggesting inherent limitations in the discriminative information present in single-cell images alone. Another limitation concerns the model's performance under extreme staining variations. While the diffusion component provides some robustness to staining differences, cases with severe staining artifacts or excessive eosinophilia still pose challenges. This indicates that incorporating explicit color normalization as a preprocessing step or during data augmentation could further improve robustness.

The current framework processes individual cell images, which may not fully leverage the contextual information available in whole slide images. Clinical diagnoses often consider the spatial distribution and relative frequencies of different cell types within a sample, information that is lost when analyzing cells in isolation. Future extensions could incorporate whole-slide level context using multiple instance learning frameworks or graph neural networks that model relationships between cells. Dataset limitations also affect model generalization. While we used multiple datasets, class imbalance remains a challenge, particularly for rare categories such as invasive carcinoma cells. Future work should explore data augmentation techniques specifically designed for medical images, such as generative models for rare class synthesis, or leverage transfer learning from related pathological domains.

V. CONCLUSION

This paper has presented DiffAFCA-Net, a novel framework for cervical precancerous cell classification that synergistically combines diffusion models with adaptive fine-grained attention mechanisms. The proposed method addresses key challenges in cervical cytology image analysis, including noise robustness, fine-grained feature capture, and generalization across diverse datasets. Extensive experiments on three cervical cell datasets demonstrate that DiffAFCA-Net outperforms state-of-the-art methods across multiple metrics, achieving accuracy improvements of 2-6% over strong baselines while maintaining better calibration and uncertainty estimation. The ablation studies confirm the complementary benefits of the diffusion framework and attention mechanisms, with each component contributing significantly to the overall performance.

ACKNOWLEDGMENT

This study was approved by the Institutional Review Board (IRB) of the Second Affiliated Hospital of Dalian Medical University under approval number KY2025-559-01. The authors declare no conflicts of interest.

REFERENCES

[1] F. Bray et al., "Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," CA: A Cancer Journal for Clinicians, vol. 74, no. 3, pp. 229–263, 2024.

[2] M. Arbyn et al., "HPV prevalence and cervical cancer screening in sub-Saharan Africa: A systematic review," Lancet Oncol., vol. 22, no. 8, pp. e315–e329, 2021.

[3] G. N. Papanicolaou, "A new procedure for staining vaginal smears," Science, vol. 95, no. 2469, pp. 438–439, 1942.

[4] J. Cuzick et al., "Overview of human papillomavirus-based and other novel options for cervical cancer screening in developed and developing countries," Vaccine, vol. 26, no. 10, pp. K29–K41, 2008.

[5] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings he IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[6] L. Zhang et al., "DeepPap: Deep convolutional networks for cervical cell classification," IEEE J. Biomed. Health Inform., vol. 21, no. 6, pp. 1633–1643, 2017.

[7] J. Hu et al., "Squeeze-and-Excitation Networks," IEEE Trans. Pattern Anal. Mach. Intell., vol. 42, no. 8, pp. 2011–2023, 2020.

[8] O. Ginsburg et al., "Breast and cervical cancer control in low and middle-income ntries: Human rights meet sound health policy," J Cancer Policy, vol. 3, pp. 14-19, 2015.

[9] Y. Yi et al., "Multi-scale window transformer for cervical cytopathology image recognition," Comput. Struct. Biotechnol. J., vol. 24, pp. 314–321, 2024.

[10] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2020.

[11] Liu Z, Lin Y, Cao Y, et al. “Swin transformer: Hierarchical vision transformer using shifted dows”[C], Proceedings of the IEEE/CVF international conference on computer vision. 2021: 12-10022.

[12] Duan Z, Xu C, Li Z, et al. “FMA-Net: Fusion of Multi-Scale Attention for Grading Cervical cancerous Lesions”[J]. Mathematics, 2024, 12(7): 958.

[13] J. Ho et al., "Denoising Diffusion Probabilistic Models," in Adv. Neural Inf. Process. Syst., vol. 33, 2020, pp. 6840–6851.

[14] Y. Yang et al., "DiffMIC: Dual-guidance diffusion network for medical image classification," in Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv. (MICCAI), 2023, pp. 95–105.

[15] A. Kazerouni et al., "Diffusion models for medical image denoising: A comparative study," Image Anal., vol. 91, p. 103050, 2023.

[16] M. M. Rahaman et al., "DeepCervix: A deep learning-based framework for the classification of cervical cells using hybrid deep feature fusion techniques," Comput. Biol. Med., vol. 136, p. 104649, 2021.

[17] M. E. Plissiti et al., "Sipakmed: A new dataset for feature and image-based classification of normal and pathological cervical cells in Pap smear images," in Proc. IEEE Int. Conf. Image Process. (ICIP), 2018, pp. 3144–3148.

[18] P. Jiang et al., "Channel spatial collaborative attention network for fine-grained classification of cervical cells," in Proc. Int. Conf. Neural Inf. Process., 2022, pp. 540–551.

[19] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in Proc. ICLR, 2015.

[20] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," Radiology, vol. 143, no. 1, pp. 29–36, 1982.

[21] Woo S, Debnath S, Hu R, et al. Convnext v2: Co-designing and scaling convnets with ked autoencoders[C]//Proceedings of the IEEE/CVF conference on computer vision and ern recognition. 2023: 16133-16142.

[22] Yue Y, Li Z. Medmamba: Vision mamba for medical image classification[J]. arXiv preprint v:2403.03849, 2024.

[23] Huo X, Sun G, Tian S, et al. HiFuse: Hierarchical multi-scale feature fusion network for medical image classification[J]. Biomedical Signal Processing and Control, 2024, 87: 105534.