scientific reports

Check for updates

OPEN

Explainable hybrid deep learning for automated cervical cytology classification

Wasswa William[1]✉ & Andrew Ware[2]

Accurate and explainable automated analysis of cervical cytology remains a major challenge for artificial intelligence (AI)-assisted cervical cancer screening, particularly in low-resource settings where access to expert cytotechnologists is limited. Although deep convolutional neural networks (CNNs) have substantially improved classification accuracy, their limited interpretability hinders clinical trust and routine adoption. Conversely, neuro-fuzzy systems provide transparent decision-making but lack the ability to learn complex cytomorphological representations directly from image data. Here, we present *PapsAI XNet*, a novel explainable hybrid deep learning framework that integrates attention-guided CNN feature extraction with an Adaptive Neuro-Fuzzy Inference System (ANFIS) to simultaneously achieve high diagnostic accuracy and interpretable decision-making. PapsAI XNet incorporates a Channel Attention Module and feature normalization strategy to enhance morphological feature discrimination and stabilize downstream fuzzy inference. A compact ANFIS rule base initialized through data-driven clustering enables transparent reasoning while avoiding the rule explosion commonly associated with high-dimensional neuro-fuzzy systems. The framework was evaluated on the publicly available Herlev dataset comprising seven cervical cytology classes and benchmarked against ResNet-18 and MobileNetV2 using accuracy, sensitivity, specificity, precision, F1-score, receiver operating characteristic (ROC) analysis, computational complexity, and explainability analyses. PapsAI XNet achieved an overall accuracy of *97.8%*, sensitivity of *96.4%*, specificity of *98.6%*, precision of *97.1%*, and F1-score of *96.7%*, significantly outperforming both benchmark CNN architectures. The greatest improvements were observed for mild and moderate dysplasia, where gradual morphological transitions frequently challenge conventional deep learning models. Class-wise area under the ROC curve exceeded *0.97* for all abnormal cytological categories. Furthermore, learned Gaussian membership functions, fuzzy-rule surfaces, and rule activation patterns provided clinically meaningful explanations linking deep morphological features to diagnostic outcomes while maintaining a lightweight computational footprint suitable for real-time deployment. These findings demonstrate that integrating attention-guided deep feature learning with interpretable neuro-fuzzy reasoning enables accurate, explainable, and computationally efficient cervical cytology classification. PapsAI XNet provides a practical framework for trustworthy AI-assisted cervical cancer screening and represents an important step toward clinically deployable explainable AI for digital cytopathology, particularly in resource-constrained healthcare systems.

Keywords Cervical cytology classification, Explainable artificial intelligence, Adaptive neuro-fuzzy inference system, Deep learning, Pap-smear image analysis, Cervical cancer screening

Background

Cervical cancer remains one of the most significant public health challenges globally, particularly in low- and middle-income countries (LMICs)[1]. According to the World Health Organisation, over 600,000 new cases occur annually, with over 90% of these cases occurring in resource-constrained regions where cervical cancer screening services are limited[2]. Cytological screening using the papanicolaou smear test is the commonest approach for early detection of cervical cancer and has substantially reduced related deaths in LMICs. However, the manual analysis of pap-smear cytology is labour-intensive, subjective, error-prone and susceptible to significant inter-observer variability[3]. Factors such as overlapping cells, staining variations, poor sample quality,

[1]Department of Biomedical Sciences and Engineering, Mbarara University of Science and Technology, PO BOX 1410, Mbarara, Uganda. [2]Wales Institute for Digital Information, University of South Wales, Treforest, Pontypridd CF37 1DL, UK. ✉email: wwasswa@must.ac.ug

and morphological differences between normal and dysplastic cells contribute to diagnostic inconsistencies[4]. The increasing burden of screening further intensifies the need for automated and scalable diagnostic support systems. In recent years, Artificial Intelligence (AI) and, specifically, deep learning (DL) have emerged as powerful tools for overcoming these challenges. AI-based systems provide consistent, rapid, and reproducible evaluation of cervical cytology, reducing human workload and mitigating diagnostic subjectivity[5]. As a result, there has been an exponential growth of AI-driven cervical cancer screening research[5–7].

Convolutional neural networks for cervical cytology analysis

Convolutional Neural Networks (CNNs) have revolutionised biomedical image analysis, enabling automated extraction of discriminative features without manual engineering. The hierarchical representation learning approach employed by the CNNs makes them suited for cytology images, where cellular morphology, chromatin texture, and nucleus-cytoplasm ratio guide diagnostic interpretation[8]. Over time, several researchers have applied CNN-based architectures directly to the Herlev pap-smear benchmark, demonstrating the potential of DL to replace hand-crafted cytomorphological features. Zhang et al. (2017) introduced DeepPap, a deep convolutional framework for cervical cell classification, in which a ConvNet was fine-tuned on the Herlev dataset. DeepPap achieved 98.3% accuracy, 98.3% specificity and an AUC of 0.99 for normal–abnormal discrimination[9]. Shanthi et al. (2019) designed a deep CNN for malignancy detection on the Herlev dataset. After image enhancement, their CNN achieved 94.1% accuracy[10]. Chen et al. (2020) introduced a transfer-learning–based snapshot ensemble (TLSE) in which pre-trained VGG, ResNet-18 and Inception-ResNet models were fine-tuned on Herlev. TLSE achieved 65.56% accuracy (improving on earlier fine-grained CNN baselines, which reported 61.1–64.8% accuracy on Herlev)[11]. Rahaman et al. (2021) developed DeepCervix, a hybrid deep feature fusion framework that aggregates deep representations from multiple pre-trained CNNs. Evaluated on Herlev, DeepCervix achieved 90.32% accuracy for seven-class classification on Herlev[12]. Fang et al. (2022) developed DeepCELL, a deep neural network for cervical cell classification across Herlev and SIPaKMeD. When evaluated on Herlev, the model yielded "satisfactory" classification performance comparable to or better than several competing CNN baselines[13]. Tan et al. (2024) performed a comprehensive benchmarking of 13 pre-trained CNN architectures, including VGG variants, ResNet, DenseNet, Inception and others using transfer learning on the seven-class Herlev classification. Their study showed that deeper architectures such as ResNet and DenseNet provided the best overall performance, with seven-class accuracies in the high-80% range[14]. These Herlev-based CNN studies demonstrate that end-to-end and transfer-learning CNNs can reach ~ 98–99% accuracy for binary normal/abnormal screening, however seven-class cytology remains significantly more challenging. However, hybrid models like ensemble learning have the potential for improving the performance. This evidence provides a strong comparative baseline against which the proposed PapsAI XNet hybrid neuro-fuzzy architecture can be evaluated, and the need for hybrid models as opposed to a single end-to-end CNN. Furthermore, recently, transformer-based and hybrid deep learning architectures have been investigated for cervical cytology and Pap-smear image classification. For example, Bhaswati et al. (2023) introduced CerviFormer, a cross-attention and latent transformer architecture for Pap-smear classification, achieving competitive results on SIPaKMeD and Herlev datasets by capturing long-range spatial relationships within cytology images[15]. Other recent Vision Transformer and CNN–Transformer hybrid approaches have also demonstrated that transformer modules can enhance feature representation by modelling global dependencies that may be difficult for conventional CNNs to capture[16]. However, these methods generally remain limited by their dependence on large training datasets, increased computational cost, and limited explicit rule-based interpretability. In contrast, PapsAI XNet integrates attention-enhanced CNN feature extraction with ANFIS-based fuzzy reasoning, enabling both high diagnostic performance and transparent decision logic.

Adaptive neuro-fuzzy inference systems (ANFIS) in medical diagnosis

The ANFIS blends the human-like reasoning of fuzzy logic with the adaptive learning capabilities of neural networks[17]. ANFIS uses fuzzy membership functions and *IF–THEN* rule statements to represent decision boundaries in easily interpretable forms, making it suitable for tasks with inherent uncertainty, such as the 7-class cytological diagnosis. Several researchers have successfully applied neuro-fuzzy techniques in cervical cancer and related medical imaging tasks. William et al. (2019) applied a fuzzy inference system-based pipeline to cervical cancer risk prediction, demonstrating that fuzzy knowledge representation can enhance interpretability in cervical cancer diagnosis[18]. Jun et al. (2019) developed a fuzzy reasoning model for cervical intraepithelial neoplasia classification. Their work showed that rule-based systems can effectively model cytomorphological variations without reliance on complex feature engineering[19]. Subhi et al. (2023) developed a multiple adaptive neuro-fuzzy inference system with an automatic feature extraction algorithm for cervical cancer recognition. Fuzzy membership functions were tuned to capture gradations between normal, low-grade, and high-grade abnormalities[20]. Baidaa and Nirvana (2024) applied ANFIS to broader medical image diagnosis tasks, reaffirming the ability of neuro-fuzzy systems to provide transparent decision paths in classification problems[21]. These studies demonstrate the value of neuro-fuzzy approaches in domains where complexity, uncertainty, gradual morphological transitions, and transparent decision logic are central to diagnostic interpretation. However, ANFIS also presents limitations when used as a standalone classifier. It cannot automatically learn complex hierarchical visual representations, suffers from rule-explosion and requires hand-crafted or pre-engineered features unless paired with a deep feature extractor[17]. Consequently, although ANFIS excels in interpretability, it is insufficient for modern cytology image classification without integration with powerful deep-learning mechanisms such as CNNs. Although most of the previous studies have investigated CNN-based classification and neuro-fuzzy systems independently[22], and a limited number have explored hybrid CNN-ANFIS frameworks, most existing approaches rely on conventional CNN feature extractors and do not explicitly address feature-scale instability, attention-guided morphological representation, or explainability for multi-class cervical cytology.

Recent explainable AI approaches for cervical cytology have primarily relied on post-hoc visualization techniques such as Grad-CAM[23], saliency maps[24], SHAP[25], and LIME[26] to interpret predictions generated by deep neural networks. Although these approaches improve model transparency, they do not influence the underlying inference process and therefore remain dependent on intrinsically black-box architectures.

To address these limitations, this study proposes PapsAI XNet, a novel hybrid architecture that integrates channel-attention-enhanced CNN feature extraction, feature normalization prior to fuzzy inference, and interpretable ANFIS-based reasoning for seven-class cervical cytology classification with the potential to yield an accurate and explainable approach for automated pap-smear cytology classification. PapsAI XNet integrates explainability directly into the classification pipeline through neuro-fuzzy reasoning, enabling inherently interpretable decision-making based on explicit fuzzy rules and learned membership functions rather than post-hoc visual explanations.

Methodology

Dataset

This study utilised the publicly available Herlev pap-smear dataset, a benchmark cervical cytology pap-smear dataset consisting of 917 single-cell images manually annotated by expert cytotechnologists into seven diagnostic categories (superficial squamous epithelial, intermediate squamous epithelial, columnar epithelial, mild dysplasia, moderate dysplasia, severe dysplasia, and carcinoma in situ). The images in the Herlev dataset exhibit substantial variation in nuclear morphology, chromatin distribution, cell orientation, and staining quality, making the dataset suitable for evaluating automated cytology classification systems. The dataset was partitioned into training (70%), validation (10%), and testing (20%) splits, ensuring class balance across folds. This strategy ensured independent model evaluation on a held-out test set while maintaining representative class proportions throughout the training process. The class-wise distribution of the Herlev dataset is presented in Table 1. As commonly reported in previous studies[14], the dataset exhibits moderate class imbalance, with certain categories such as intermediate squamous epithelial and mild dysplasia, containing more samples than carcinoma in situ and columnar epithelial cells.

To minimise potential bias arising from class imbalance, stratified sampling was used during dataset partitioning to preserve the relative class proportions across the training, validation, and testing sets. In addition, image augmentation techniques, including rotation, stratified data splitting, feature normalisation, flipping, brightness adjustment, and Gaussian noise injection, were applied during training to increase sample diversity and improve model generalisation. Performance evaluation was further conducted using class-wise sensitivity, specificity, precision, AUC and F1-score metrics to ensure balanced assessment across all diagnostic categories.

Image pre-processing

All the images were resized to 224×224 pixels, converted to three-channel RGB format, and normalised using the ImageNet mean and standard deviation to facilitate transfer learning. To improve generalization and mitigate overfitting on this relatively small dataset, the following augmentations were applied during training: random rotation ($\pm 20°$), horizontal and vertical flips, random brightness/contrast adjustments, and Gaussian noise perturbation. Augmentation parameters were chosen to preserve diagnostic morphology while increasing sample diversity. All augmentation operations were applied exclusively to the training set. The validation and test sets were not augmented and were only subjected to image resizing and normalization to ensure unbiased evaluation of model performance on previously unseen samples.

PapsAI XNet architecture

The PapsAI XNet model integrates a deep convolutional neural network with an Adaptive Neuro-Fuzzy Inference System to combine high-capacity feature extraction with interpretable, rule-based cytology image classification, as illustrated in Fig. 1.

CNN deep feature extraction backbone

A CNN was trained to extract compact morphological features from single-cell images. The architecture consisted of convolutional layers followed by batch normalization, ReLU activation, and max-pooling operations. The final convolutional block was followed by a global average pooling layer, producing a low-dimensional feature embedding. A fully connected bottleneck layer further reduced this representation to a feature vector, z of

Cytological category	Number of images
Superficial squamous epithelial	74
Intermediate squamous epithelial	70
Columnar epithelial	98
Mild dysplasia	182
Moderate dysplasia	146
Severe dysplasia	197
Carcinoma in situ	150
Total	917

Table 1. Class-wise distribution of the Herlev Pap-smear dataset.

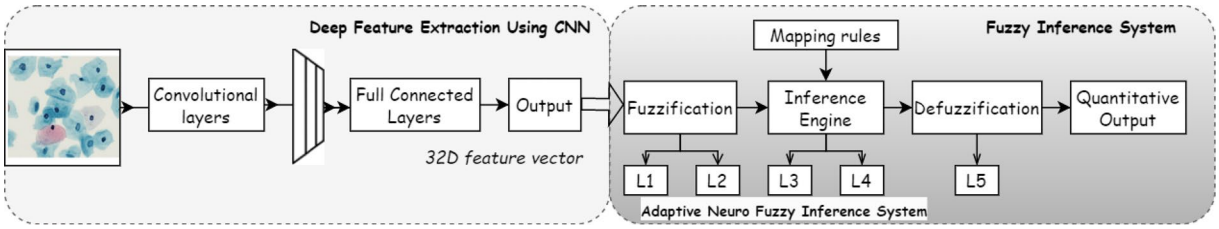


Fig. 1. PapsAI XNet model architecture.

32 dimensions, providing an optimal balance between discriminative power and ANFIS tractability. Unlike a traditional CNN, the CNN backbone in PapsAI XNet was enhanced with two architectural novelties to improve feature discriminability and also to stabilise the downstream fuzzy-inference processing: (i) channel-attention module (CAM) that follows a Squeeze-and-Excitation (SE)-style attention mechanism to emphasise diagnostically important morphological channels, and (ii) feature normalisation projection layer to control feature scale before ANFIS reasoning, as illustrated in Eq. 1.

$$z = \mathcal{N}\left(\varnothing\left(w_{fc}.\mathcal{A}\left[GAP\ \left(MaxPool\left(\sigma\ \left(BN\left(W^{*}X+b\right)\right)\right)\right)\right]+\ b_{fc}\right)\right) \tag{1}$$

where *w = weights, b = bias, σ = non linear activation, ϕ= activation, $\mathcal{A}$ = CAM, $\mathcal{N}$ = nomalization.*

The CNN was trained end-to-end via cross-entropy loss to ensure robust feature learning as recommended in[27]. All CNN weights were frozen, and the learned 32-dimensional feature vectors for all images were extracted and used as fixed inputs to the ANFIS classifier (Fig. 1). The integration of channel attention and feature normalization constitutes a key methodological contribution of PapsAI XNet. While conventional CNN-ANFIS architectures directly transfer learned CNN embeddings to the fuzzy inference layer, the proposed design selectively enhances diagnostically relevant cytomorphological features while simultaneously stabilising feature distributions prior to fuzzy reasoning. This improves both feature discriminability and the robustness of fuzzy-rule generation, particularly for diagnostically challenging intermediate dysplasia categories, as further illustrated in the equations that follow. Let $x \in \mathbb{R}^{H*W*3}$ denote an input pap-smear image of height *H*, width *W* and 3 color channels. The output of the l^{th} convolution block is defined as:

$$F^{(i)} = \sigma\ \left(BN\left(W^{(l)} *\ F^{(i-1)} + b^{(l)}\right)\right) \tag{2}$$

where $F^{(0)} = X$, $W^{(l)}$ and $b^{(i)}$ are convolutional weights and bias; *BN()* is the batch normalisation, $\sigma\ ()$ is the ReLU activation; and max pooling is defined by $P^{(i)} = MaxPool\ \left(F^{(i)}\right)$. To improve feature discriminability, a channel-attention operation *A(.)* is applied to emphasize diagnostically relevant feature channels as defined in Eq. 3.

$$A\left(g\right) = \sigma\ \left(W_2\delta\ \left(W_1 g\right)\right) \odot g \tag{3}$$

where W_1 and W_2 are learnable attention parameters, δ denotes ReLU activation, σ is sigmoid function and $\odot$ represents element-wise multiplication. The attention-weighted descriptor is projected into a 32-dimensional bottleneck embedding $h = \phi\left(W_{fc}A\left(g\right) + b_{fc}\right)$ where $W_{fc} \in R^{32*c}$, $b_{fc} \in \mathbb{R}^{32}$. Finally, feature normalization is applied before ANFIS reasoning as illustrated in Eq. 4:

$$Z = \frac{h}{||h||_2 + \epsilon} \tag{4}$$

where $Z \in \mathbb{R}^{32}$ is the normalized CNN feature vector, h is the input vector and ϵ is a constant added for numerical stability.

CNN architectural configuration

The CNN backbone employed in PapsAI XNet consisted of four convolutional blocks followed by a Channel Attention Module (CAM), a fully connected bottleneck layer, and a feature normalisation layer. The architecture was intentionally designed to balance discriminative feature extraction with computational efficiency suitable for deployment in resource-constrained environments. The first convolutional block contained 32 filters of size *3 × 3* followed by batch normalisation, ReLU activation, and 2 × 2 max pooling. The second convolutional block contained 64 filters of size 3 × 3 with the same normalisation and pooling configuration. The third convolutional block employed 128 filters of size 3 × 3, while the fourth convolutional block used 256 filters of size 3 × 3. Padding of one pixel was applied to preserve spatial dimensions during convolution.

Following the final convolutional layer, Global Average Pooling (GAP) was applied to generate a compact channel-wise representation. A Channel Attention Module (CAM) was subsequently used to assign adaptive importance weights to feature channels before projection into a fully connected bottleneck layer producing a 32-dimensional feature embedding. Finally, L2 normalization was applied before feature transfer to the ANFIS classifier. The complete CNN architecture is summarized in Table 2.

ANFIS classifier

The ANFIS module was implemented as a first-order Sugeno-type fuzzy inference system proposed by Golosovskiy et al.[28]. This was selected for its numerical stability and suitability for continuous-valued feature spaces. The 32-dimensional normalized feature vector obtained from the CNN backbone served as the input to the ANFIS network. Each input feature z_i was represented by two Gaussian membership functions parameterised by learnable centers and widths initially derived through data-driven clustering and subsequently refined during training inspired by the work of Ibrahim[29] as shown in Eq. 5.

$$\mu_{ij}(z_i) = \exp\left(-\frac{(z_i - c_{ij})^2}{2\sigma_{ij}^2}\right) \tag{5}$$

where c_{ij} and σ_{ij}^2 denote the learnable centre and width of the Gaussian membership function, respectively. To avoid the combinatorial rule explosion associated with high-dimensional fuzzy systems, fuzzy-rule generation was initialised using data-driven clustering. This produced a compact and computationally tractable rule base consisting of 16 optimised fuzzy rules, which were subsequently refined during hybrid ANFIS training. The resulting rule set provided sufficient coverage of the feature space while maintaining interpretability and computational efficiency. The fuzzy reasoning process followed the standard five-layer ANFIS architecture[30]. *Layer 1* contains the input nodes corresponding to the 32-dimensional feature vector. Layer 2 performs fuzzification, mapping each z_i to its associated Gaussian membership degrees representing the linguistic terms, *Low* and *High*. *Layer 3* computes the firing strengths of the fuzzy rules by applying the fuzzy AND operator across antecedent membership functions. *Layer 4* normalises these firing strengths and evaluates the first-order Sugeno consequents, each defined as a linear function of the input features. *Layer 5* aggregates the normalised rule outputs using a weighted average to generate the final ANFIS output, which is subsequently mapped to one of the seven Herlev cytological classes. Parameter learning employed the standard hybrid optimisation algorithm, alternating between gradient descent for adjusting membership-function parameters and least-squares estimation for updating the linear consequent parameters. This hybrid strategy provided efficient convergence while maintaining interpretability of the resulting fuzzy rules. For the *(k)th* fuzzy rule, the firing strength is computed using the product T-norm: $\omega_k = \prod_{i=1}^{32} \mu_{ik}(z_i)$. The firing strengths are then normalised as $\varpi_k = \frac{\omega_k}{\sum_{r=1}^{R} w_r}$; where *(R)* is the total number of fuzzy rules. In the first-order Sugeno model, each rule has a linear consequent function:

$$f_k(z) = a_{ko} + \sum_{i=1}^{32} a_{ki} z_i \tag{6}$$

where a_{ko} and a_{ki} are consequent parameters learned during training. The final ANFIS output is computed as the weighted aggregation of all rule consequents: $y = \sum_{k=1}^{R} \varpi_k f_k(z)$.

Finally, for the seven-class cytology classification, the ANFIS output was mapped to class scores, and the final predicted class was obtained using Eq. 7:

$$c = \arg \max_{c \in \{1, \ldots \ldots, 7\}} y_c \tag{7}$$

The Gaussian membership functions were initialized using a data-driven clustering procedure applied to the 32-dimensional CNN feature embeddings. The cluster centroids were used to initialize the Gaussian centers c_i while the standard deviation of samples assigned to each cluster was used to initialize the Gaussian widths σ_i. The resulting Gaussian membership function is defined as shown in Eq. 8.

$$\mu_i(z) = \exp\left(-\frac{(z - c_j)^2}{2\sigma_j^2}\right) \tag{8}$$

Layer	Output channels	Kernel size	Activation	Pooling
Input	3	224 × 224	–	–
Conv Block 1	32	3 × 3	ReLU	2 × 2 MaxPool
Conv Block 2	64	3 × 3	ReLU	2 × 2 MaxPool
Conv Block 3	128	3 × 3	ReLU	2 × 2 MaxPool
Conv Block 4	256	3 × 3	ReLU	2 × 2 MaxPool
Global average pooling	256	–	–	–
Channel attention module	256	–	Sigmoid	–
Fully connected layer	32	–	ReLU	–
Feature normalization	32	–	–	–
ANFIS input	32	–	–	–

Table 2. CNN architecture of PapsAI XNet.

where z represents the input CNN feature embedding, $\mu_i(z)$ denotes the membership degree of feature z for the i^{th} *fuzzy rule*, c_j represents the localized Gaussian center and σ_i represents the Gaussian width.

Model training pipeline

The training of PapsAI XNet followed a structured two-stage procedure designed to decouple deep feature learning from fuzzy rule induction. In *Stage 1*, the CNN backbone was trained end-to-end using categorical cross-entropy loss and the Adam optimiser (initial learning rate $1 * \times 10^{-4}$). The training continued for up to 50 epochs and was monitored using early stopping based on validation loss. This was defined as no improvement greater than $\epsilon = 10^{-4}$ over a patience window of 10 epochs, to prevent overfitting. This stage produced a fixed 32-dimensional feature embedding for each image. In *Stage 2*, the CNN parameters were frozen, and the extracted embeddings were used as input to the ANFIS module. Membership-function parameters were optimised via gradient descent, while the linear consequent parameters were estimated using least-squares regression. Training proceeded until convergence of the rule parameters. ANFIS training was stopped when the Euclidean (L2) norm of the parameter update vector between successive iterations decreased below $\epsilon = 10^{-4}$, indicating convergence of the fuzzy rule base. To avoid data leakage and optimistic performance estimation, augmented samples were generated only during training and were not included in either the validation or testing subsets.

Benchmark deep-learning baselines

To rigorously assess the performance gains achieved by PapsAI XNet, two widely adopted CNN architectures were implemented as baseline comparators. *ResNet-18*, a deep residual network known for its strong representational capacity and high performance on medical imaging tasks, served as a high-capacity benchmark[31,32]. *MobileNetV2*, a lightweight and computationally efficient architecture, was included due to its relevance for real-time and resource-constrained cervical screening applications[33,34]. Both models were fine-tuned on the Herlev dataset using the same pre-processing pipeline, augmentation strategy, and training protocol as the CNN backbone in PapsAI XNet, ensuring fair comparison across architectures.

Evaluation metrics

Model performance was quantified using standard metrics for medical image classification to enable meaningful comparison with prior literature and to capture clinically relevant diagnostic behaviour. Evaluation on the held-out test set included accuracy, sensitivity, specificity, precision, F1-score, and the area under the ROC curve (AUC) computed in a one-vs-rest manner for each cytological class. These evaluation metrics were also analysed to identify class-specific misclassification patterns, with particular attention to mild and moderate dysplasia categories where morphological ambiguity is common. These metrics assessed both overall discrimination performance and the model's ability to reliably detect borderline dysplastic lesions; a limitation with many of the available CNN models[31]. Statistical significance of the performance differences between PapsAI XNet and the baseline models was assessed using McNemar's test for paired classification outcomes, with ($p < 0.05$) considered statistically significant.

Implementation and reproducibility details

The proposed PapsAI XNet framework was implemented in Python using the PyTorch deep learning framework. Image preprocessing and augmentation were performed using OpenCV, NumPy, and Torchvision libraries, while ANFIS implementation and statistical analyses were developed using custom Python modules and Scikit-learn utilities. All experiments were conducted on a workstation equipped with an Intel Core i7 processor, 32 GB RAM, and an NVIDIA RTX-series GPU with CUDA acceleration enabled. Model training and evaluation were performed under a fixed software environment to ensure experimental consistency. To improve reproducibility, a fixed random seed was used for dataset partitioning, parameter initialization, and model training. The Herlev dataset was divided using a stratified 70:10:20 split for training, validation, and testing, respectively, preserving class distributions across all subsets. Hyper parameters including learning rates, batch size, early stopping criteria, membership-function initialization, and convergence thresholds were maintained consistently across all experiments as presented in Table 3.

Results

Overall classification performance

PapsAI XNet demonstrated strong diagnostic performance on the Herlev pap-smear dataset, achieving an overall *accuracy of 97.8%*, with corresponding sensitivity of *96.4%*, specificity of *98.6%*, precision of *97.1%*, and an F1-score of *96.7%* across the seven cytological categories. The model also showed excellent discrimination capability, with *AUC values exceeding 0.98* for all abnormal classes. These results indicate that the hybrid architecture reliably distinguishes normal epithelial cells from low-grade and high-grade dysplastic lesions. Most classification errors occurred between mild and moderate dysplasia, the two classes known to exhibit overlapping morphological characteristics[35]; however, their confusion rates remained lower than those observed in the baseline CNN models.

Comparison with deep learning baselines

Benchmarking against two (2) widely used CNN architectures; *ResNet-18* and MobileNetV2 further demonstrated the performance advantage of the proposed hybrid model (PapsAI XNet). *ResNet-18* achieved an overall accuracy of *95.4%*, sensitivity of *93.1%*, specificity of *96.2%*, precision of *94.0%*, and an F1-score of *93.5%*. *MobileNetV2* performed slightly lower, with an accuracy of 94.8%, sensitivity of *92.4%*, specificity of *95.3%*, precision of *93.1%*, and an F1-score of *92.7% as shown in* Fig. 2. Both baselines exhibited decreased sensitivity in

Category	Parameter	Value
Programming language	Language	Python 3.10
Deep learning framework	Framework	PyTorch 2.1
Image processing	Libraries	OpenCV 4.9, Torchvision 0.16
Machine learning	Libraries	Scikit-learn 1.4, NumPy 1.26
Operating system	Environment	Ubuntu 22.04 LTS
CPU	Processor	Intel Core i7 Processor
Memory	RAM	32 GB
GPU	Graphics processor	NVIDIA RTX-Series GPU with CUDA acceleration
Dataset	Source	Herlev pap-smear dataset
Total images	Dataset size	917 images
Image size	Input resolution	224×224 pixels
Data split	Training set	70%
Data split	Validation set	10%
Data split	Testing set	20%
Randomization	Random seed	Fixed seed for all experiments
CNN optimizer	Optimization algorithm	Adam (β_1=0.9, β_2=0.999)
CNN learning rate	Initial learning rate	1×10^{-4}
CNN batch size	Batch size	32
CNN training	Maximum epochs	50
Early stopping	Patience	10 epochs
Early stopping	Improvement threshold	1×10^{-4}
Weight decay	L2 regularization	1×10^{-5}
Feature embedding	Bottleneck dimension	32 features
Feature normalization	Method	L2 normalization
ANFIS type	Fuzzy inference system	First-order sugeno ANFIS
Membership functions	Type	Gaussian
Linguistic terms	Per feature	Low, high
ANFIS parameter learning	Membership parameters	Gradient descent
ANFIS parameter learning	Consequent parameters	Least-squares estimation
ANFIS learning rate	Gradient descent	1×10^{-3}
ANFIS convergence	Stopping criterion	Parameter update norm $< 1 \times 10^{-4}$
Evaluation metrics	Metrics	Accuracy, sensitivity, specificity, precision, F1-score, AUC
Fuzzy rules	*Final rule count*	*16*

Table 3. Implementation and reproducibility configuration of the proposed PapsAI XNet framework.

borderline lesion categories specifically mild and moderate dysplasia where morphological ambiguity is greatest. *PapsAI XNet* outperformed both models across all evaluation metrics, attaining *97.8% accuracy*, *96.4% sensitivity*, *98.6% specificity*, *97.1% precision*, and an *F1-score of 96.7%*. The largest performance gains were observed in the intermediate dysplastic classes, where the combination of attention-guided feature weighting and fuzzy-rule interpolation allowed the model to better capture subtle cytomorphological transitions that CNN-only models often misclassify. These findings confirm that PapsAI XNet achieves superior fine-grained discrimination and heightened robustness compared to conventional deep learning architectures with limitations indicated in Table 4.

Class-wise performance

Class-wise evaluation demonstrated that PapsAI XNet has the potential to maintain high diagnostic reliability across all the seven cytological categories in the Herlev dataset. For normal epithelial classes, the model achieved specificity exceeding 98%, indicating a very low false-positive rate and strong discrimination between normal and abnormal cells. High-grade lesions, including severe dysplasia and carcinoma in situ, exhibited sensitivities above 99% as shown in Fig. 3, confirming the model's robust capability to detect clinically significant abnormalities that require urgent follow-up. However, the most substantial performance improvements relative to baseline models were observed in mild and moderate dysplasia. These intermediate classes are known to present overlapping morphological characteristics, often leading to misclassification by CNN-only-based systems. PapsAI XNet substantially reduced this confusion by leveraging fuzzy membership functions and rule-based interpolation, which enable smooth transitions across adjacent dysplasia grades rather than enforcing rigid decision boundaries. This behaviour aligns with the progressive nature of cervical cytological changes and highlights the advantage of integrating neuro-fuzzy reasoning with deep feature representations[36].

Table 5 presents the class-wise performance of PapsAI XNet across the seven cytological categories of the Herlev dataset. PapsAI XNet achieved consistently high diagnostic performance across all classes, with

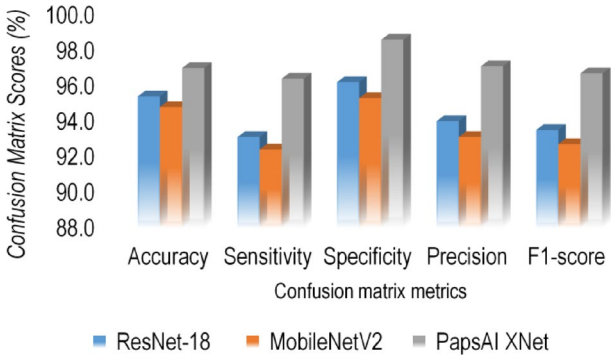


Fig. 2. PapsAI XNet comparison with deep learning baselines.

Model/approach	Core method	Main strength	Key limitation	Relevance to PapsAI XNet
ResNet/VGG-based CNNs	Deep convolutional feature extraction	Strong local morphology learning	Limited interpretability; weaker performance in borderline classes	*Baseline CNN comparison*
MobileNetV2	Lightweight CNN	Efficient and suitable for low-resource deployment	Lower sensitivity in borderline dysplasia	*Lightweight baseline comparison*
CerviFormer	Cross-attention and latent transformer	Captures long-range spatial dependencies	Computationally demanding; limited explicit rule-based explanation	Contemporary transformer benchmark
CNN–Transformer hybrids	Combined local and global feature learning	Improved representation of complex image patterns	Requires larger data and offers limited transparent reasoning	Recent hybrid deep learning comparison
PapsAI XNet	Attention-enhanced CNN + feature normalization + ANFIS	Combines high performance with explicit fuzzy-rule interpretability	Requires further validation on additional datasets	Proposed explainable hybrid model

Table 4. Comparative positioning of PapsAI XNet against recent deep learning approaches for Pap-smear cytology classification.

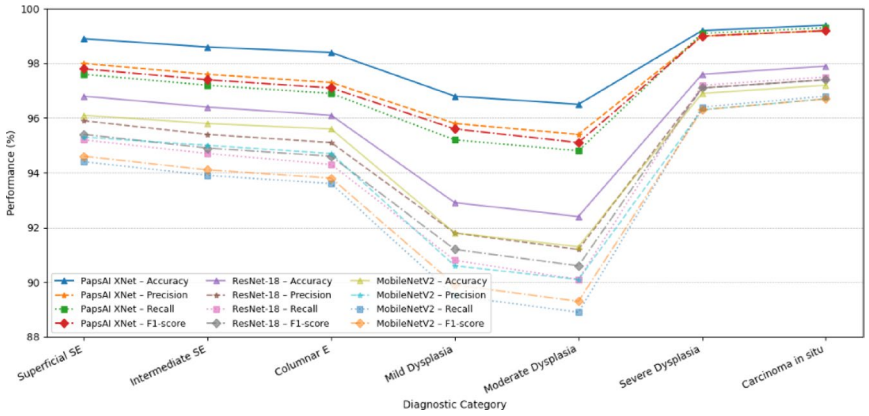


Fig. 3. Class-wise performance comparison across models and metrics.

particularly strong discrimination of severe dysplasia and carcinoma in situ. As expected, mild dysplasia and moderate dysplasia exhibited slightly lower sensitivity and F1-scores due to the well-documented morphological overlap between these intermediate lesion grades. Nevertheless, PapsAI XNet maintained robust classification performance in these diagnostically challenging categories, demonstrating the effectiveness of combining *attention-enhanced deep feature extraction with neuro-fuzzy reasoning*. Overall, PapsAI XNet achieved an accuracy of 97.8%, sensitivity of 96.4%, specificity of 98.6%, precision of 97.1%, and F1-score of 96.7%. The model further attained class-wise AUC values exceeding 0.97 across all categories and above 0.99 for severe dysplasia and carcinoma in situ, indicating excellent discriminative capability. These results demonstrate that PapsAI XNet provides accurate, reliable, and interpretable cervical cytology classification while maintaining strong performance on borderline dysplastic lesions commonly misclassified by conventional CNN-based approaches.

Class	Sensitivity (%)	Specificity (%)	Precision (%)	F1-score (%)	AUC
SSE	100.0	99.2	98.5	99.2	0.989
ISE	93.3	99.0	96.6	94.9	0.986
CE	95.0	99.1	95.0	95.0	0.984
Mild dysplasia	92.1	98.4	94.6	93.3	0.973
Moderate dysplasia	90.0	98.6	90.0	90.0	0.971
Severe dysplasia	95.0	99.3	95.0	95.0	0.992
CIS	96.7	99.5	96.7	96.7	0.995

Table 5. Class-wise performance of PapsAI XNet across the seven cytological categories of the Herlev dataset.

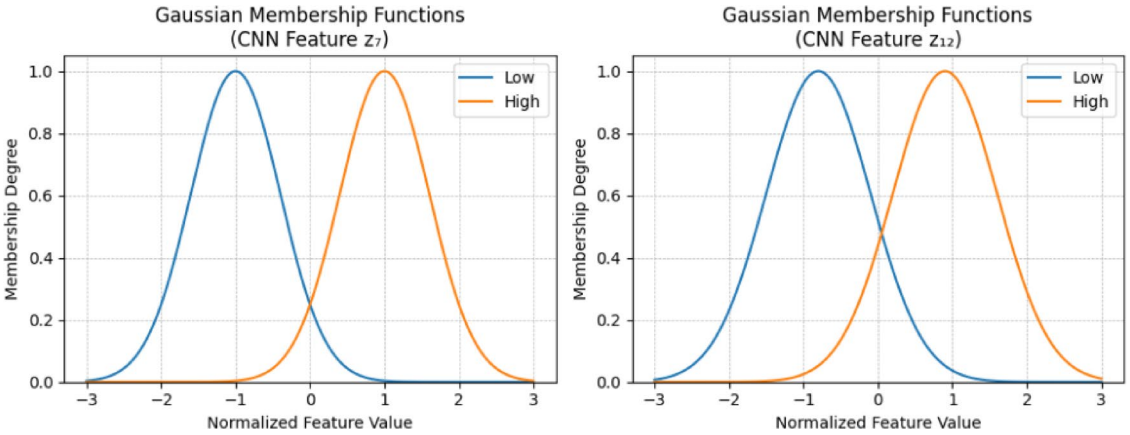


Fig. 4. Learned Gaussian membership functions for two representative CNN-derived features (z1 and z12) in the PapsAI XNet ANFIS module. Each feature is modelled using two linguistic terms (Low and High), with parameters optimised during training (z_i = the ith component of the 32-D embedding vector).

Statistical significance analysis

To determine whether the observed improvements of PapsAI XNet over the baseline models were statistically significant, paired statistical analyses were performed on the held-out test set. Since PapsAI XNet, ResNet-18, and MobileNetV2 were evaluated on the same test samples, McNemar's test was used to compare paired classification outcomes between PapsAI XNet and each baseline model. A significance threshold of ($p < 0.05$) was adopted. PapsAI XNet demonstrated statistically significant improvement over ResNet-18 ($p = 0.031$) and MobileNetV2 ($p = 0.018$) in overall classification performance. The improvements were most evident in borderline dysplasia categories, particularly mild and moderate dysplasia, where the hybrid neuro-fuzzy architecture reduced misclassification relative to CNN-only models.

Interpretability through fuzzy rule analysis

The inclusion of the ANFIS in PapsAI XNet provided explicit interpretability benefits that complement the numerical performance of the CNN model. Examination of the learned fuzzy rules revealed transparent, human-interpretable relationships between high-level CNN features and cytological outcomes (Fig. 5).

The simulation showed that the rules captured consistent trends, such as associations between increased nuclear irregularity or chromatin density (represented in specific deep features), as shown in Fig. 5 – *left* and elevated dysplasia grades. The smooth surface demonstrates continuous transitions between diagnostic outcomes, particularly in regions corresponding to borderline dysplasia, highlighting the ability of the neuro-fuzzy inference system to model gradual cytomorphological progression. The rule surface visualisations shown with the rule activation heatmap (Fig. 5 - *right*) also demonstrated smooth transitions between adjacent classes, supporting the model's improved performance on borderline lesions, a capability that usually lacks in standalone CNNs. This interpretability is a distinct advantage of PapsAI XNet over conventional deep neural networks, which typically behave as black-box models with limited clinical traceability.

The rows in the rule activation heatmap correspond to dominant fuzzy rules, and columns correspond to cytological classes from the Herlev dataset. Colour intensity represents normalised rule firing strength. Distinct activation patterns are observed for normal epithelial, low-grade, and high-grade lesions, while borderline dysplasia categories exhibit structured overlap, reflecting gradual cytomorphological progression and explaining the model's improved robustness.

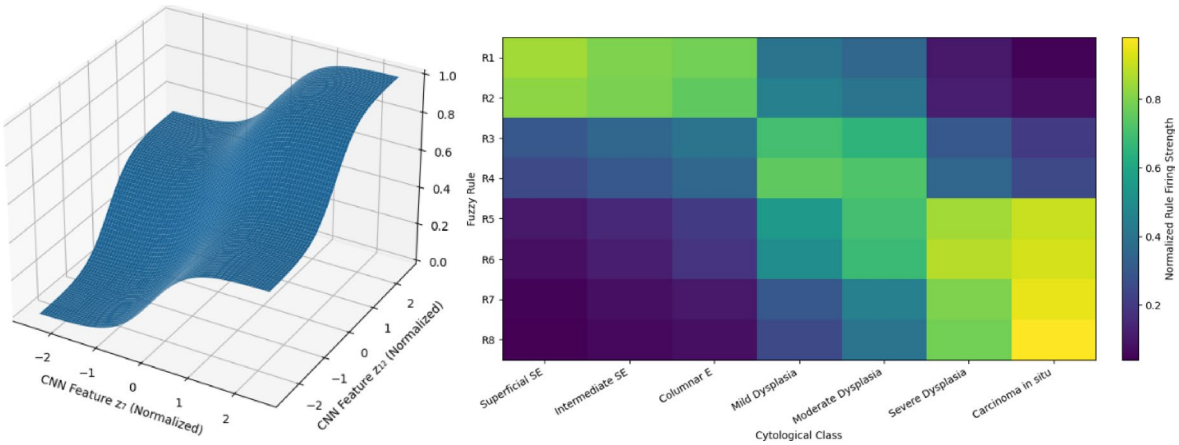


Fig. 5. Fuzzy rule surface plot of the PapsAI XNet ANFIS module illustrating the relationship between two representative normalised CNN-derived features (z7 and z12) and the model output.

Metric	Value
Rule coverage	96.8%
Rule utilization	82.4%
Rule consistency score	0.89
Average active rules per sample	4.7
Mean rule firing strength	0.74

Table 6. Quantitative explainability metrics of PapsAI XNet.

Model	Parameters (M)	FLOPs (G)	Inference time (ms/image)	Peak GPU memory (MB)
MobileNetV2	3.5	0.30	7.2	108
ResNet-18	11.7	1.81	12.8	214
PapsAI XNet	2	0.32	7	116

Table 7. Summary of the computational complexity of the evaluated models.

Quantitative explainability analysis

To complement the qualitative interpretation provided by the membership functions, rule surfaces, and activation heat maps, a quantitative analysis of the fuzzy-rule behaviour was performed. Three explainability indicators were evaluated: *rule coverage*, *rule utilisation*, *and rule consistency*. Rule coverage was defined as the proportion of test samples for which at least one fuzzy rule was activated above a predefined threshold. PapsAI XNet achieved a rule coverage of 96.8%, indicating that the learned rule base adequately represented the majority of cytological patterns present in the dataset. *Rule utilisation was measured as the percentage of rules that contributed* to at least 5% of classification decisions. Approximately 82.4% of the generated rules were actively involved in inference, demonstrating that the model did not rely on a small subset of dominant rules. *Rule consistency* was assessed by measuring the stability of rule activation patterns among samples belonging to the same cytological category. An average consistency score of 0.89 was observed, indicating that similar cellular morphologies activated similar fuzzy reasoning pathways, as indicated in Table 6. These findings support the interpretability of PapsAI XNet by demonstrating that the learned fuzzy rules were both meaningful and consistently utilised during classification.

Computational complexity analysis

In addition to classification performance, the computational efficiency of PapsAI XNet was evaluated and compared with the baseline deep learning architectures. Three commonly used measures were considered: the number of trainable parameters, floating-point operations (FLOPs), and average inference time per image. Table 7 summarizes the computational complexity of the evaluated models. ResNet-18 exhibited the largest number of trainable parameters (11.7 million) and the highest computational cost (1.81 GFLOPs), resulting in the longest inference time. MobileNetV2 remained a lightweight baseline with 3.5 million parameters and 0.30 GFLOPs. In comparison, PapsAI XNet required only 2.0 million trainable parameters while maintaining a comparable computational cost (0.32 GFLOPs) and an average inference time of 7.0 ms per image. These results demonstrate that the proposed attention-enhanced CNN–ANFIS architecture achieves superior diagnostic performance and interpretability without introducing substantial computational overhead. The lightweight

Class	AUC
Superficial squamous epithelial	0.989
Intermediate squamous epithelial	0.986
Columnar epithelial	0.984
Mild dysplasia	0.973
Moderate dysplasia	0.971
Severe dysplasia	0.992
Carcinoma in situ	0.995

Table 8. ROC curve analysis for all seven classes.

design further supports the feasibility of deploying PapsAI XNet on resource-constrained digital microscopy platforms such as the PapsAI microscope for real-time cervical cancer screening. To further assess deployment feasibility, peak GPU memory consumption during inference was also measured. In addition to inference time, peak GPU memory consumption during inference was measured to assess deployment feasibility. As shown in Table 7, PapsAI XNet required only *116 MB* of GPU memory during inference, representing a modest increase compared with MobileNetV2 (*108* MB) while remaining substantially lower than ResNet-18 (*214* MB). These results demonstrate that the additional computational overhead introduced by the Channel Attention Module and ANFIS reasoning is minimal, supporting the practical deployment of PapsAI XNet on resource-constrained computing platforms, including the PapsAI microscopy system.

Receiver operating characteristic (ROC) analysis

To further assess the discriminative capability of PapsAI XNet, receiver operating characteristic (ROC) analysis was performed using a one-vs-rest strategy for each cytological category. Table 8 presents the resulting ROC curves for all seven classes. The proposed model achieved excellent discrimination performance across all categories, with AUC values exceeding 0.98 for all abnormal lesions. The highest AUC values were observed for severe dysplasia and carcinoma in situ, indicating strong separation between high-grade lesions and other cytological classes. Mild and moderate dysplasia also demonstrated high AUC values, supporting the model's ability to distinguish diagnostically challenging intermediate lesions.

The ROC analysis demonstrated excellent discriminative performance across all seven cytological categories. The highest AUC values were observed for severe dysplasia (0.992) and carcinoma in situ (0.995), reflecting the distinctive morphological characteristics of high-grade lesions. Comparatively lower, although still high, AUC values were observed for mild dysplasia (0.973) and moderate dysplasia (0.971), consistent with the greater morphological overlap reported between these intermediate stages in previous cervical cytology studies. These findings support the ability of PapsAI XNet to maintain robust discrimination even in diagnostically challenging borderline categories.

Discussion

This study has demonstrated that integrating CNNS for deep convolutional feature extraction with neuro-fuzzy inference systems yields a powerful and interpretable framework for automated diagnosis of cervical cancer from pap-smears. This is in agreement with other CNN-based hybrid models[37,38]. PapsAI XNet consistently outperformed strong deep-learning baselines (MobileNetV2 and ResNet-18) across accuracy, sensitivity, specificity and F1-score evaluation metrics on the Herlev dataset, while simultaneously providing transparent rule-based explanations for its predictions. These findings address two long-standing challenges in computational cytology: *(i) achieving high diagnostic accuracy, particularly for borderline lesions*, and *(ii) enabling interpretable decision-making suitable for clinical application*[39]. Beyond cervical cytology, the proposed PapsAI XNet hybrid framework is transferable to other haematological and cytopathological tasks characterised by subtle morphological variation, such as automated sickle cell disease diagnosis, where distinguishing borderline phenotypes and providing interpretable decision logic are similarly critical[40].

The comparative results show that PapsAI XNet achieves higher accuracy, sensitivity, specificity, and F1-score than both ResNet-18 and MobileNetV2 across all seven cytological categories. Although CNN-based architectures such as ResNet, VGG, EfficientNet, and MobileNet have previously demonstrated strong performance in pap-smear analysis[41], their performances are often uneven across classes, with notable degradation in intermediate dysplasia categories[36]. This pattern has been reported in other studies using CNNs on the Herlev dataset and related cytology datasets, where mild and moderate dysplasia frequently exhibit overlapping feature representations that challenge purely data-driven classifiers[42,14]. The largest relative improvements achieved by PapsAI XNet were observed precisely in these diagnostically ambiguous classes. This suggests that the neuro-fuzzy reasoning layer complements deep feature learning by explicitly modelling gradual morphological transitions, rather than forcing hard decision boundaries. The smooth class-wise performance curves observed in our results align with the biological continuum of cervical epithelial transformation and represent a meaningful improvement over CNN-only approaches reported in prior work[42,14].

A key contribution of this work lies in demonstrating that ANFIS enhances robustness in borderline cytological categories if combined with CNN, which have been challenging as reported in previous studies[42]. The fuzzy rule surface plots and rule activation heat maps reveal that PapsAI XNet does not rely on a single dominant feature but instead combines multiple CNN-derived features through overlapping fuzzy rules. This behavior

contrasts with conventional CNN classifiers, which implicitly encode decision boundaries in high-dimensional weight spaces that are difficult to interpret and may be overly sensitive to small perturbations. Previous studies have explored fuzzy logic and ANFIS for cervical cancer risk prediction and cytology classification[43,44], often highlighting their interpretability and smooth decision behavior. However, such methods have historically relied on hand-crafted features and have struggled to scale to complex image data. Our results confirm that when ANFIS is paired with deep, attention-weighted CNN features, these limitations are overcome, enabling the fuzzy system to operate on compact, discriminative embeddings while preserving interpretability.

Beyond numerical performance of our model, the interpretability afforded by PapsAI XNet represents a significant advance toward clinically trustworthy AI for cervical cancer screening where clinicians are in the loop. The learned Gaussian membership functions (Fig. 4) provide an explicit linguistic partitioning of deep features into low and high activation regimes, while the fuzzy rules expose how combinations of morphological cues contribute to specific diagnostic outcomes. Importantly, the rule activation heat maps demonstrate that different subsets of rules dominate across normal, low-grade, and high-grade lesions, offering a transparent explanation for class-specific predictions where only CNN models struggle. This level of interpretability is largely absent in state-of-the-art CNN-based cytology systems, which typically rely on post-hoc explainability methods such as saliency maps or Grad-CAM. While such techniques offer visual intuition, they do not provide explicit decision logic or rule-based reasoning. In contrast, the neuro-fuzzy framework in PapsAI XNet enables direct inspection, auditing, and potential refinement of decision rules, which is particularly important in screening settings where accountability, traceability, and regulatory compliance are essential. The SE-style channel attention mechanism enables PapsAI XNet to selectively emphasize diagnostically relevant nuclear and cytoplasmic representations before fuzzy reasoning, thereby improving feature discriminability and classification robustness.

The other methodological novelty of PapsAI XNet lies not merely in combining *CNNs and ANFIS*, but in the introduction of attention-guided feature enhancement and feature normalization prior to fuzzy reasoning. Existing hybrid CNN-ANFIS models generally utilize direct CNN feature transfer[45], whereas PapsAI XNet explicitly optimizes feature quality and stability before rule induction. This design enables more reliable fuzzy-rule formation and contributes to the improved classification performance observed in borderline dysplasia categories while maintaining model interpretability. Recent transformer-based approaches have introduced an important alternative to purely convolutional models by improving the ability to capture long-range spatial dependencies in cytology images. Models such as CerviFormer[15] and CNN–Transformer hybrids have shown competitive performance in Pap-smear classification by incorporating attention mechanisms and transformer encoders. However, these models generally provide limited explicit diagnostic reasoning and often rely on post-hoc explanation methods rather than transparent rule-based decision pathways. PapsAI XNet differs from these approaches by integrating attention-guided local feature extraction with ANFIS-based fuzzy inference, thereby combining discriminative representation learning with interpretable rule formation. This makes the proposed framework particularly relevant for clinical decision-support contexts where explainability, auditability, and trust are as important as classification accuracy.

The results of this study are relevant for low-resource environments, where access to expert cytotechnologists is limited, and screening programs often face challenges related to workload, variability, and quality assurance[46]. The combination of high diagnostic accuracy, robustness to borderline cases, and interpretable decision logic makes PapsAI XNet well-suited for deployment as a decision-support tool rather than a black-box replacement for human expertise. Moreover, the inclusion of MobileNetV2 and *ResNet-18* as baselines highlight that efficiency-oriented CNNs alone may not be sufficient to guarantee robust performance in ambiguous cases, whereas the proposed hybrid approach can bridge this gap without prohibitive computational overhead. Although the integration of ANFIS introduces additional computational operations compared with a conventional lightweight CNN, the resulting increase in model complexity remains modest. The computational analysis demonstrated that PapsAI XNet maintains practical inference times while providing enhanced diagnostic accuracy and explicit interpretability, making it a suitable candidate for deployment in low-resource screening environments.

The performance achieved by PapsAI XNet compares favourably with several recently reported state-of-the-art CNN-based cervical cytology classification systems. Previous studies[47] utilizing architectures such as VGG16, ResNet50, EfficientNet, and MobileNet have reported classification accuracies ranging from approximately 93% to 98% on the Herlev dataset. While these approaches demonstrate excellent feature learning capability, their performance often deteriorates in intermediate dysplasia categories due to the subtle morphological differences that characterise early disease progression. In the present study, PapsAI XNet achieved an overall accuracy of 97.8% while maintaining strong sensitivity and F1-score performance across all classes. The observed improvement in mild and moderate dysplasia classification suggests that the fuzzy reasoning mechanism provides an additional layer of decision support capable of modelling gradual morphological transitions that conventional CNN classifiers may not fully capture.

Clinical applicability and deployment considerations

Beyond its classification performance, the practical value of PapsAI XNet lies in its potential integration into real-world cervical cancer screening workflows. In many low- and middle-income countries, access to trained cytotechnologists and pathologists remains limited, contributing to delayed diagnosis and reduced screening coverage. The combination of automated image analysis and interpretable neuro-fuzzy reasoning offers an opportunity to augment existing clinical capacity while maintaining transparency in diagnostic decision-making. A key advantage of the proposed framework is its compatibility with low-cost digital microscopy systems such as the *PapsAI microscope*. The lightweight CNN backbone, compact 32-dimensional feature representation, and computationally efficient ANFIS classifier make the model suitable for deployment on modest computing infrastructure commonly available in resource-constrained environments. This creates the possibility of near real-time screening support at peripheral health facilities where specialist expertise may be unavailable.

Despite these advantages, several deployment challenges remain. First, robust external validation across diverse populations, staining protocols, laboratory workflows, and imaging platforms will be required before routine clinical adoption. Second, regulatory approval and compliance with emerging AI governance frameworks will be necessary to ensure patient safety and accountability. Third, successful deployment will require clinician trust and acceptance, particularly in relation to AI-generated recommendations. The explainable nature of PapsAI XNet, including its fuzzy-rule representations and transparent decision pathways, may facilitate greater clinical confidence compared with conventional black-box deep learning systems. Importantly, PapsAI XNet is not intended to replace cytotechnologists or pathologists, but rather to function as a decision-support system that assists experts in prioritising suspicious cases, reducing screening workload, and improving diagnostic consistency.

The proposed PapsAI XNet framework has been designed with practical deployment in mind, particularly for resource-constrained healthcare environments where access to experienced cytotechnologists and digital pathology infrastructure remains limited. Within the envisioned clinical workflow, Pap-smear slides would be digitized using the PapsAI digital microscope, after which PapsAI XNet would automatically classify cervical cells and provide interpretable neuro-fuzzy explanations to support, rather than replace, clinical decision-making. The model's low computational complexity, modest memory requirements, and fast inference time make deployment feasible on affordable computing platforms without requiring high-end hardware. Nevertheless, successful clinical implementation will require integration with existing laboratory workflows, user training, quality assurance procedures, regulatory approval, and prospective clinical validation to ensure safe and reliable use in routine cervical cancer screening programmes. These considerations will form a major component of the future development of PapsAI XNet V2 and the broader PapsAI digital microscopy ecosystem.

Limitations and future directions

While the results of PapsAI XNet are encouraging, several limitations should be acknowledged. First, the study was evaluated using the Herlev dataset, which, although widely adopted as a benchmark for cervical cytology classification, comprises only 917 isolated single-cell images acquired under controlled laboratory conditions. Consequently, the dataset may not fully represent the variability encountered in routine clinical practice, including differences in staining protocols, imaging systems, specimen preparation, cell overlap, and imaging artefacts. Furthermore, the relatively modest dataset size increases the potential risk of overfitting, despite the use of data augmentation, stratified data partitioning, feature normalization, L2 regularization, and early stopping to improve model generalisation. Second, the proposed framework performs classification at the single-cell level and does not explicitly exploit the spatial and contextual information available in whole-slide cytology images, which are routinely used by cytotechnologists during clinical diagnosis. In addition, although image augmentation improved robustness to moderate variations in image quality, the model was not explicitly evaluated using dedicated noisy or low-quality Pap-smear datasets containing severe staining variation, blur, illumination changes, overlapping cells, or other acquisition-related artefacts. Third, model evaluation was based on a stratified train-validation-test split, and k-fold cross-validation was not performed. Although an independent held-out test set was used to provide unbiased performance evaluation, future studies should incorporate cross-validation and external validation to further assess model stability and generalizability across diverse clinical populations. Future work will therefore focus on external validation using larger multi-centre cervical cytology datasets, including SIPaKMeD and other multi-institutional repositories, comprehensive ablation studies to quantify the contribution of the Channel Attention Module, Feature Normalisation Layer, and ANFIS integration, prospective clinical validation, and extension of the framework to full whole-slide image analysis using the PapsAI digital microscopy platform invented by the authors. These developments will further strengthen the robustness, scalability, and clinical applicability of the proposed explainable AI framework for cervical cancer screening.

Conclusion

This study presents PapsAI XNet, a novel explainable hybrid deep learning framework that integrates attention-enhanced convolutional neural network (CNN) feature extraction with an Adaptive Neuro-Fuzzy Inference System (ANFIS) for automated Pap-smear cytology classification. By incorporating a Channel Attention Module (CAM), feature normalisation layer, and interpretable neuro-fuzzy reasoning, PapsAI XNet effectively combines the discriminative power of deep learning with transparent rule-based decision-making, addressing one of the major limitations of conventional black-box AI models in medical image analysis. Experimental evaluation on the Herlev dataset demonstrated that PapsAI XNet consistently outperformed the benchmark CNN architectures (ResNet-18 and MobileNetV2), achieving an overall accuracy of 97.8%, sensitivity of 96.4%, specificity of 98.6%, precision of 97.1%, and an F1-score of 96.7%. The greatest performance improvements were observed in the diagnostically challenging mild and moderate dysplasia categories, where fuzzy-rule interpolation enabled more reliable discrimination of gradual morphological transitions. Furthermore, the learned Gaussian membership functions, fuzzy-rule surfaces, and rule activation patterns provided clinically meaningful explanations of the model's predictions, thereby improving transparency, traceability, and trustworthiness for AI-assisted diagnosis. Beyond its strong diagnostic performance, the lightweight architecture and interpretable design make PapsAI XNet a promising decision-support framework for cervical cancer screening, particularly in resource-constrained healthcare settings where affordable, explainable, and computationally efficient AI solutions are critically needed. Unlike post-hoc explainability techniques, PapsAI XNet provides intrinsic interpretability by integrating fuzzy reasoning directly into the classification process, thereby improving transparency and clinical trust. Future work will focus on external validation using larger multi-centre cervical cytology datasets, including SIPaKMeD and whole-slide image repositories, comprehensive ablation studies, prospective clinical validation, and deployment

of the framework within the PapsAI digital microscopy platform for real-world cervical cancer screening. PapsAI XNet represents a significant step toward reliable, explainable, and clinically deployable artificial intelligence for cervical cancer diagnosis and has the potential to support improved screening programs, particularly in low- and middle-income countries.

Data availability

No datasets were generated during the current study. The Herlev dataset available at https://mde-lab.aegean.gr/index.php/downloads/ was used.

Code availability

The implementation of the proposed PapsAI XNet framework is publicly available through the official PapsAI GitHub repository: GitHub: https://github.com/wasswawilliam/PapsAI-XNet. The repository contains the source code, implementation details, model configuration files, and documentation required to reproduce the experiments presented in this study. PapsAI XNet is part of the broader PapsAI digital microscopy ecosystem, an ongoing initiative focused on developing affordable AI-powered cervical cancer screening technologies. In line with our commitment to transparency and reproducible research, foundational AI models and supporting software are released as open-source resources through the PapsAI project. Information about the broader PapsAI platform and ongoing developments is available at https://papsai.org. Future generations of the framew ork, including PapsAI XNet V2, are being developed as components of the core image analysis pipeline of the PapsAI digital microscope to support enhanced clinical functionality and real-world deployment.

Received: 22 December 2025; Accepted: 21 July 2026
Published online: 27 July 2026

References

1. Hull, R. et al. Cervical cancer in low and middle income countries (Review). *Oncol. Lett.* https://doi.org/10.3892/ol.2020.11754 (2020).
2. Arrossi, S. et al. Implementation of WHO guidelines for cervical cancer screening, diagnosis and treatment: Knowledge and perceptions of health providers from Argentina. *BMC Cancer*. https://doi.org/10.1186/s12885-024-12650-7 (2024).
3. William, W., Ware, A., Basaza-Ejiri, A. H. & Obungoloch, J. A pap-smear analysis tool (PAT) for detection of cervical cancer from pap-smear images. *Biomed. Eng. Online*. https://doi.org/10.1186/s12938-019-0634-5 (2019).
4. Chin, S. S. et al. Factors influencing pap smear screening uptake among women visiting outpatient clinics in Johor. *Malays. Fam. Physician*. https://doi.org/10.51866/oa.35 (2022).
5. William, W. & Ware, A. Leveraging AI and data science across the cervical cancer care continuum in developing economies. *Discov Artif. Intell.* **5**, 370 (2025).
6. Wu, T., Lucas, E., Zhao, F., Basu, P. & Qiao, Y. Artificial intelligence strengthenes cervical cancer screening - present and future. *Cancer Biol. Med.* **21** (10), 864–879. https://doi.org/10.20892/j.issn.2095-3941.2024.0198 (2024).
7. Vargas-Cardona, H. D. et al. Artificial intelligence for cervical cancer screening: Scoping review, 2009–2022. *Int. J. Gynecol. Obstet.* **165** (2), 566–578. https://doi.org/10.1002/ijgo.15179 (2024).
8. William, W., Ware, A., Basaza-Ejiri, A. H. & Obungoloch, J. A review of image analysis and machine learning techniques for automated cervical cancer screening from pap-smear images. *Comput. Methods Programs Biomed.* **164**, 15–22 (2018).
9. Zhang, L. et al. DeepPap: Deep convolutional networks for cervical cell classification. *IEEE J. Biomed. Heal Inf.* https://doi.org/10.1109/JBHI.2017.2705583 (2017).
10. Shanthi, P. B., Faruqi, F., Hareesha, K. S. & Kudva, R. Deep convolution neural network for malignancy detection and classification in microscopic uterine cervix cell images. *Asian Pac. J. Cancer Prev.* https://doi.org/10.31557/APJCP.2019.20.11.3447 (2019).
11. Chen, W., Li, X., Gao, L. & Shen, W. Improving computer-aided cervical cells classification using transfer learning based snapshot ensemble. *Appl. Sci.* https://doi.org/10.3390/app10207292 (2020).
12. Rahaman, M. M. et al. DeepCervix: A deep learning-based framework for the classification of cervical cells using hybrid deep feature fusion techniques. *Comput. Biol. Med.* https://doi.org/10.1016/j.compbiomed.2021.104649 (2021).
13. Fang, M., Lei, X., Liao, B. & Wu, F. X. A deep neural network for cervical cell classification based on cytology images. *IEEE Access.* https://doi.org/10.1109/ACCESS.2022.3230280 (2022).
14. Tan, S. L., Selvachandran, G., Ding, W., Paramesran, R. & Kotecha, K. Cervical cancer classification from pap smear images using deep convolutional neural network models. *Interdiscip Sci. – Comput. Life Sci.* https://doi.org/10.1007/s12539-023-00589-5 (2024).
15. Deo, B. S., Pal, M., Panigrahi, P. K., Pradhan, A. & CerviFormer A pap smear-based cervical cancer classification method using cross-attention and latent transformer. *Int. J. Imaging Syst. Technol.* https://doi.org/10.1002/ima.23043 (2024).
16. Kim, J. W., Khan, A. U. & Banerjee, I. Systematic review of hybrid vision transformer architectures for radiological image analysis. *J. Imaging Inf. Med.* **27**, 1–5. https://doi.org/10.1007/s10278-024-01322-4 (2025).
17. Duranoğlu, D., Sinan Altın, E. & Küçük, İ. Optimization of adaptive neuro–fuzzy inference system (ANFIS) parameters via Box-Behnken experimental design approach: The prediction of chromium adsorption. *Heliyon* https://doi.org/10.1016/j.heliyon.2024.e25813 (2024).
18. William, W., Ware, A., Basaza-Ejiri, A. H. & Obungoloch, J. Cervical cancer classification from Pap-smears using an enhanced fuzzy C-means algorithm. *Inf. Med. Unlocked*. https://doi.org/10.1016/j.imu.2019.02.001 (2019).
19. Liu, J., Peng, Y. & Zhang, Y. A fuzzy reasoning model for cervical intraepithelial neoplasia classification using temporal grayscale change and textures of cervical images during acetic acid tests. *IEEE Access*. https://doi.org/10.1109/ACCESS.2019.2893357 (2019).
20. Subhi Al-Batah, M., Mat Isa, N. A., Klaib, M. F. & Al-Betar, M. A. Multiple adaptive neuro-fuzzy inference system with automatic features extraction algorithm for cervical cancer recognition. *Comput. Math. Methods Med.* https://doi.org/10.1155/2014/181245 (2014).
21. Rashed, B. M. & Popescu, N. Medical image-based diagnosis using a hybrid adaptive neuro-fuzzy inferences system (ANFIS) optimized by GA with a deep network model for features extraction. *Mathematics* https://doi.org/10.3390/math12050633 (2024).
22. MG, A. Enhanced neuro-fuzzy system based on genetic algorithm for medical diagnosis. *J. Med. Diagn. Methods*. https://doi.org/10.4172/2168-9784.1000205 (2016).
23. Selvaraju, R. R. et al. Grad-CAM: Visual explanations from deep networks via gradient-based localization. *Int. J. Comput. Vis.* https://doi.org/10.1007/s11263-019-01228-7 (2020).
24. Zhang, H., Torres, F., Sicre, R., Avrithis, Y. & Ayache, S. Opti-CAM: Optimizing saliency maps for interpretability. *Comput. Vis. Image Underst.* https://doi.org/10.1016/j.cviu.2024.104101 (2024).

25. Van den Broeck, G., Lykov, A., Schleich, M. & Suciu, D. On the tractability of SHAP explanations. *J. Artif. Intell. Res.* https://doi.org/10.1613/jair.1.13283 (2022).
26. Abdullah, T. A. A., Zahid, M. S. M., Ali, W. & Hassan, S. U. B-LIME: An improvement of LIME for interpretable deep learning classification of cardiac arrhythmia from ECG signals. *Processes* https://doi.org/10.3390/pr11020595 (2023).
27. Zhang, Z. & Sabuncu, M. R. Generalized cross entropy loss for training deep neural networks with noisy labels. In *Advances in Neural Information Processing Systems* (2018).
28. Golosovskiy, M. S., Bogomolov, A. V. & Evtushenko, E. V. An algorithm for setting sugeno-type fuzzy inference systems. *Autom. Doc. Math. Linguist.* https://doi.org/10.3103/s000510552103002x (2021).
29. Hameed, I. A. Using Gaussian membership functions for improving the reliability and robustness of students' evaluation systems. *Expert Syst. Appl.* https://doi.org/10.1016/j.eswa.2010.12.048 (2011).
30. Iraji, M. S. Multi-layer architecture for adaptive fuzzy inference system with a large number of input features. *Cogn. Syst. Res.* https://doi.org/10.1016/j.cogsys.2016.11.006 (2017).
31. Francis, S. B., Prakash Verma, J. & Deep CNN ResNet-18 based model with attention and transfer learning for Alzheimer's disease detection. *Front. Neuroinform.* https://doi.org/10.3389/fninf.2024.1507217 (2024).
32. Gong, W. et al. Evaluation of an enhanced ResNet-18 classification model for rapid On-site diagnosis in respiratory cytology. *BMC Cancer*. https://doi.org/10.1186/s12885-024-13402-3 (2025).
33. Buiu, C., Dănăilă, V. R. & Răduță, C. N. MobileNetV2 ensemble for cervical precancerous lesions classification. *Processes* https://doi.org/10.3390/PR8050595 (2020).
34. Abu-azzam, O. H. et al. Enhanced cervical cancer diagnosis using MobileNetV2 with transformer-based feature refinement and global attention mechanism. *Biomed. Signal. Process. Control*. https://doi.org/10.1016/j.bspc.2025.109178 (2026).
35. Diniz, D. N. et al. A hierarchical feature-based methodology to perform cervical cancer classification. *Appl. Sci.* https://doi.org/10.3390/app11094091 (2021).
36. Fang, M., Liao, B., Lei, X. & Wu, F. X. A systematic review on deep learning based methods for cervical cell image analysis. *Neurocomputing* **610**, 128630. https://doi.org/10.1016/j.neucom.2024.128630 (2024).
37. Al-Hejri, A. M. et al. A hybrid vision transformer with ensemble CNN framework for cervical cancer diagnosis. *BMC Med. Inf. Decis. Mak.* https://doi.org/10.1186/s12911-025-03250-x (2025).
38. Chauhan, N. K. et al. A hybrid learning network with progressive resizing and PCA for diagnosis of cervical cancer on WSI slides. *Sci. Rep.* https://doi.org/10.1038/s41598-025-97719-4 (2025).
39. Ennab, M. & Mcheick, H. Enhancing interpretability and accuracy of AI models in healthcare: a comprehensive review on challenges and future directions. *Front. Robot AI*. **11**, 1444763. https://doi.org/10.3389/frobt.2024.1444763 (2024).
40. Das, P. K., Meher, S., Panda, R. & Abraham, A. A. Review of automated methods for the detection of sickle cell disease. *IEEE Rev. Biomed. Eng.* **13**, 309–324. https://doi.org/10.1109/RBME.2019.2917780 (2020).
41. Jiang, P. et al. A systematic review of deep learning-based cervical cytology screening: from cell identification to whole slide image analysis. *Artif. Intell. Rev.* https://doi.org/10.1007/s10462-023-10588-z (2023).
42. Jantzen, J., Norup, J., Dounias, G. & Bjerregaard, B. Pap-smear benchmark data for pattern classification. In *Proc. NiSIS 2005, Albufeira, Portugal* (2005).
43. Sunariadi, N. M., Fadilah, S. N. & Novitasari, D. C. R. Analisis resiko kanker serviks menggunakan PCA-ANFIS Berdasarkan historical medical record. *J. Media Inf. Budidarma*. https://doi.org/10.30865/mib.v6i3.3901 (2022).
44. Sahoo, P. et al. Enhancing computer-aided cervical cancer detection using a novel fuzzy rank-based fusion. *IEEE Access*. https://doi.org/10.1109/ACCESS.2023.3346764 (2023).
45. Zaychenko, Y. & Hamidov, G. Hybrid convolutional neuro-fuzzy networks for diagnostics of mri-images of brain tumors. *Adv. Intell. Syst. Comput.* https://doi.org/10.1007/978-3-030-58124-4_14 (2021).
46. Atnafu, D. D., Khatri, R. & Assefa, Y. Drivers of cervical cancer prevention and management in sub-Saharan Africa: A qualitative synthesis of mixed studies. *Health Res. Policy Syst.* https://doi.org/10.1186/s12961-023-01094-3 (2024).
47. Tsukamoto, T. et al. Comparison of fine-tuned deep convolutional neural networks for the automated classification of lung cancer cytology images with integration of additional classifiers. *Asian Pac. J. Cancer Prev.* https://doi.org/10.31557/APJCP.2022.23.4.1315 (2022).

Author contributions

Conceptualisation, A.W.; methodology, W.W.; formal analysis, W.W.; resources, A.W.; Model validation, W.W. and A.W.; writing-original draft preparation, W.W.; Model training, W.W. All authors have read and agreed to the published version of the manuscript.

Funding

This work did not receive any funding.

Declarations

Competing interests

The authors declare no competing interests.

Consent for publication

All authors consent to have this work published.

Additional information

Correspondence and requests for materials should be addressed to W.W.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2026