Cancer Cell International

RESEARCH

Open Access

Joint segmentation and classification of cervical cells using U-Net and CNN ensemble for early detection of cervical cancer

Akinbowale Nathaniel Babatunde[1], Oluwasegun Julius Aroba[2,6], Abidemi Emmanuel Adeniyi[3,4*], Salman Olatunji Isiaka[5], Tosho A. Abdulrahman[5], Akeem Femi Kadri[1], Shuaib Babatunde Mohammed[1] and Michael Rudolph[6]

Abstract

Cervical cancer remains a leading cause of mortality among women, particularly in resource-limited settings where cytological screening is hindered by manual analysis challenges. This study presents a joint segmentation and classification framework integrating U-Net and a CNN ensemble to automate and enhance the accuracy of cervical cell analysis. The U-Net architecture is employed to delineate nuclei and cytoplasmic boundaries from Pap smear images, leveraging skip connections to preserve spatial resolution in complex cellular environments. Following segmentation, an ensemble classifier combining ResNet-50, VGG-16, and InceptionV3 is utilised to differentiate cervical cell types with high precision. The proposed method was evaluated on benchmark datasets (APACS-23, Cx22, SIPaKMeD) using stratified five-fold cross-validation and external validation cohorts. Experimental results achieved a mean Dice coefficient of 0.934 for segmentation and a classification accuracy of 94.6%, outperforming single-model baselines by 5–8% in F1-score and AUC metrics. Grad-CAM visualisations demonstrated that the model's focus corresponded to diagnostically relevant regions, reinforcing clinical interpretability. Ablation studies confirmed that the joint learning architecture substantially enhanced feature localisation and classification robustness. This research contributes a deployable, interpretable AI-based pipeline suitable for real-time cervical cancer screening, offering potential deployment in both centralised laboratories and low-resource healthcare environments. Future directions include expanding dataset diversity, integrating transformer-based models, and developing lightweight implementations for edge devices.

Keywords Cervical Cancer Screening, Joint Segmentation and Classification, U-Net, CNN Ensemble, Medical Image Analysis, Explainable AI

*Correspondence:
Abidemi Emmanuel Adeniyi
abidemi.adeniyi@bowen.edu.ng
[1]Department of Computer Science, Kwara State University, Malete, Kwara State, Nigeria
[2]Department of Operations and Quality Management, Durban University of Technology, Durban, South Africa
[3]Department of Information Systems, Faculty of Computer Science and Engineering, Obafemi Awolowo University, Ile-Ife, Osun State., Nigeria
[4]Chitkara University Institute of Engineering and Technology, Chitkara University, Jansla, Punjab, India
[5]Department of Computer Science, Kwara State Polytechnic, Ilorin, Kwara State, Nigeria
[6]Centre for Ecological Intelligence, Faculty of Engineering & the Built Environment, University of Johannesburg, Johannesburg, South Africa

© The Author(s) 2026. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, 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 you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. 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-nc-nd/4.0/.

Introduction

Cervical cancer remains a leading cause of cancer-related mortality among women worldwide, particularly in low- and middle-income countries where access to regular cytological screening is limited. Conventional Pap smear screening relies heavily on manual microscopic examination by trained cytopathologists, a process that is time-consuming, labour-intensive, and prone to inter-observer variability and diagnostic inconsistency (Adler et. al., 2025; Pacal, 2025; Sang'udi et al., 2025). These challenges are exacerbated by the intrinsic complexity of cervical cytology images, which often exhibit staining variability, overlapping cells, irregular nuclear morphology, and poorly defined cytoplasmic boundaries (Garbach et al., 2025; Gruber et al., 2025). As screening volumes increase, these limitations pose a significant barrier to timely and accurate early detection (Li, Liu, & Xie, 2023).

Recent advances in deep learning have substantially transformed medical image analysis, enabling automated and robust feature learning directly from raw image data. In cervical cytology, convolutional neural networks (CNNs) have demonstrated strong performance in both segmentation and classification tasks, outperforming traditional image processing and classical machine learning approaches ([27]; Adler et al., 2024; Shaik et al., 2025). Encoder–decoder architectures such as U-Net have become the de facto standard for biomedical image segmentation due to their ability to preserve fine-grained spatial details through skip connections, even when training data are limited (Zhang et al., 2023). Accurate segmentation of nuclei and cytoplasm is particularly critical in cervical cancer screening, as diagnostic decisions rely heavily on morphological cues such as nuclear size, shape, chromatin texture, and nucleus-to-cytoplasm ratio (Babatunde et al., 2025a; Babatunde et al., 2025b).

Parallel to segmentation advances, CNN-based classification models have shown high diagnostic accuracy in distinguishing cervical cell types. Architectures such as ResNet-50, VGG-16, and InceptionV3 have been widely adopted due to their hierarchical and multi-scale feature extraction capabilities (Kadri et al., 2025a; Kadri et al., 2025b) [25]. However, single-network classifiers often suffer from limited generalization, especially when applied across heterogeneous datasets. Ensemble learning strategies mitigate this limitation by combining complementary models, thereby reducing variance and improving robustness [22]. Ensemble CNNs have consistently reported improved F1-score and AUC metrics on benchmark cervical cytology datasets such as SIPaKMeD and Herlev (Heidari et al., 2024; Kamzayeva, 2025).

Despite these advances, many existing studies treat segmentation and classification as independent or loosely coupled stages, resulting in fragmented pipelines that limit interpretability and clinical trust. More recent joint-learning frameworks attempt to integrate segmentation and classification, but several gaps remain. These include limited evaluation across diverse datasets, insufficient coupling between spatial localisation and diagnostic inference, and a lack of explainability mechanisms that clearly link predictions to clinically meaningful cellular structures (Zhang et al., 2023; Shaik et al., 2025). Addressing these gaps is essential for transitioning deep learning systems from experimental settings to real-world cervical cancer screening workflows.

This study proposes an interpretable, end-to-end joint segmentation and classification framework for automated cervical cell analysis. The framework integrates a U-Net–based segmentation module with a soft-voting CNN ensemble comprising ResNet-50, VGG-16, and InceptionV3. Segmentation outputs are used not merely as a preprocessing step but as spatial priors that directly guide classification, ensuring that diagnostic decisions are grounded in anatomically relevant regions (Lin et al., 2018; Wubineh, Rusiecki, Halawa, 2024; Tomaziu-Todosia et al., 2025). The framework is evaluated across three complementary public datasets, APACS-23, Cx22, and SIPaKMeD, covering pixel-level segmentation, overlapping-cell scenarios, and multi-class diagnostic classification (Yang et al., 2024; Wang et al., 2025). Furthermore, segmentation-guided Grad-CAM visualizations are incorporated to enhance interpretability, allowing clinicians to verify that model attention aligns with diagnostically relevant nuclei and cytoplasmic regions ([27]; Shaik et al., 2025).

The main contributions of this work are threefold. First, the study introduces a tightly coupled joint-learning architecture that explicitly aligns segmentation and ensemble classification through joint optimization, improving both localization accuracy and diagnostic robustness. Second, comprehensive multi-dataset validation is conducted to demonstrate generalizability across diverse cytological imaging conditions, addressing a key limitation of many prior studies that rely on single-dataset evaluation (Heidari et al., 2024). Third, the framework embeds clinically grounded explainability by linking segmentation outputs with Grad-CAM–based visual interpretation, thereby enhancing transparency and supporting clinician trust. Collectively, these contributions advance the state of the art by delivering a unified, interpretable, and generalizable deep learning pipeline suitable for deployment in real-world cervical cancer screening environments.

Literature review

Recent progress in automated cervical cancer screening has been driven by advances in machine learning and deep learning applied to cytological image analysis. Traditional Pap smear interpretation relies on expert visual

assessment of cellular morphology, a process that is time-consuming and prone to inter- and intra-observer variability, particularly in resource-constrained settings (Chauhan, Singh, & Kumar, 2023; Chauhan, Singh, & Mehta, 2025). Early computational approaches sought to mitigate these challenges through handcrafted feature extraction combined with conventional machine learning classifiers. Such methods focused on morphological descriptors, texture statistics, and colour features, achieving moderate success but exhibiting limited robustness to staining variability, overlapping cells, and heterogeneous acquisition conditions [12, 24]. A pivotal transition toward data-driven representation learning occurred with the adoption of convolutional neural networks (CNNs), which enabled end-to-end optimisation directly from raw cytology images. Comparative studies by Chauhan and Singh [5] provided early empirical evidence that deep learning architectures consistently outperform conventional machine learning pipelines in cervical cell classification, particularly in complex multi-class scenarios. Subsequent investigations reinforced these findings, demonstrating that CNNs are more effective at capturing subtle morphological variations associated with dysplastic progression than manually engineered features [25, 26]. However, classification-focused CNN models frequently lack explicit spatial grounding, limiting interpretability and clinical trust (Chauhan, Singh, & Verma, 2023).

Parallel to classification research, segmentation-based deep learning methods emerged to address the need for precise localisation of diagnostically relevant structures such as nuclei and cytoplasm. Encoder–decoder architectures, particularly U-Net and its variants, became the dominant paradigm for biomedical image segmentation due to their ability to preserve spatial detail through skip connections ([27]; Zhang et al., 2023). High segmentation accuracy has been reported across diverse cervical cytology datasets, including challenging overlapping-cell scenarios (Heidari et al., 2024; [23]). Nevertheless, most segmentation-oriented studies remain diagnostically incomplete, as localisation outputs are not explicitly integrated into downstream classification or clinical decision-making processes. To bridge the gap between localisation and diagnosis, hybrid learning strategies have been increasingly explored. These approaches combine handcrafted descriptors with deep features or integrate multiple deep representations to exploit complementary information. For example, Chauhan and Singh , [7] demonstrated that selective feature-based machine learning classifiers can improve stability under constrained data conditions, while Chauhan and Singh [6] analysed the influence of CNN channel variation on feature discrimination in cervical cytology images. More recent work introduced hybrid deep feature concatenation networks that fuse representations from multiple CNN backbones, yielding improved classification accuracy (Chauhan & Singh, 2023; [28]). Although effective, these hybrid models typically adopt modular pipelines in which segmentation, feature extraction, and classification are loosely coupled, increasing susceptibility to error propagation and limiting end-to-end optimisation.

Transfer learning and fine-tuning strategies have further advanced cervical cytology classification by adapting large-scale pretrained networks to domain-specific data [18, 25]. Chauhan and Singh (2023) reported that fine-tuned CNNs outperform training-from-scratch approaches, particularly when annotated cytology datasets are limited. Building on this, Chauhan and Singh (2025) proposed hybrid learning frameworks incorporating dimensionality reduction techniques such as principal component analysis (PCA) and progressive resizing to improve computational efficiency and convergence stability. Despite these methodological refinements, most existing frameworks continue to treat segmentation as an optional preprocessing step rather than an integral component of diagnostic inference (Zhang, Zhu, & Luo, 2025) . Recent studies have begun exploring joint or multi-task learning paradigms that integrate segmentation and classification within a unified framework. Joint learning has been shown to enhance feature localisation and reduce overfitting by enforcing shared representations across tasks (Chen et al., 2024; Shaik et al., 2025). However, current joint-learning implementations exhibit several limitations. Many are evaluated on a single dataset, limiting evidence of generalisability [14, 18]. Others prioritise performance metrics without embedding explainability mechanisms that clearly associate predictions with clinically meaningful cellular regions [22, 29]. As a result, the translation of such systems into real-world screening workflows remains constrained.

Explainable artificial intelligence (XAI) has therefore become an essential research direction in cervical cancer imaging. Visual explanation techniques such as Grad-CAM enable post hoc inspection of CNN decision regions, supporting clinician validation of automated predictions [27, 34]. While promising, explainability is often applied only at the classification stage, without explicit linkage to segmentation outputs or morphological structures. This disconnect reduces the clinical interpretability of hybrid and joint-learning systems. The limitations and methodological gaps identified across prior studies are synthesised in Table 1, which systematically contrasts existing cervical cytology frameworks in terms of modelling strategy, algorithmic design, and unresolved challenges.

Figure 1 further maps these gaps, namely suboptimal segmentation accuracy in complex cellular environments,

Table 1 Summary of Representative Cervical Cancer Imaging Studies

Author (Year)	Model	Algorithm	Contribution	Gap / Limitation
Rasheed et al., [27]	Improved U-Net	Encoder–decoder CNN with skip connections	Achieved high-accuracy nucleus segmentation (Dice ≈ 0.93) on Pap smear images	Focused solely on segmentation; no diagnostic classification or clinical interpretability
Heidari et al., (2024)	DenseNet + U-Net	Instance-level segmentation	Robust segmentation of overlapping cervical cells (IoU ≈ 0.88) on Cx22 dataset	Segmentation-only pipeline; lacks diagnostic inference
Karadag et al., [18]	ResNet, VGG	CNN-based classification	Demonstrated strong classification accuracy on SIPaKMeD dataset	Class imbalance affects robustness; no spatial localisation
Qian et al., [25]	CNN Ensemble	Voting–stacking ensemble learning	Achieved high AUC (> 0.96) for cervical cell classification	Classification-only framework; segmentation not integrated
Jin et al., [14]	Attention-based CNN	Feature fusion with attention mechanisms	Improved diagnostic accuracy using attention-guided learning	Attention lacks explicit anatomical grounding via segmentation
Liu et al., [22]	Swin Transformer + CNN	Hybrid transformer–CNN ensemble	Enhanced classification performance across datasets	High computational cost; limited deployment feasibility
Khozaimi, [19]	Dense-UNet	Metaheuristic optimisation (SMO)	Achieved very high segmentation accuracy (Dice ≈ 0.96)	No classification or end-to-end diagnostic integration
Mosquera-Zamudio et al., [23]	DL-based segmentation model	Multi-scale CNN segmentation	Scalable nucleus segmentation on whole-slide images	Does not support diagnostic classification
Shaik et al., (2025)	U-Net + CNN Ensemble	Joint segmentation and classification	Demonstrated benefits of joint learning with Grad-CAM interpretability	Limited external validation and dataset diversity
Proposed Framework (2025)	U-Net + CNN Ensemble	Joint learning with soft-voting and Grad-CAM	Unified segmentation, classification, and explainability with multi-dataset validation	Higher computational cost than single models; scalability to larger cohorts required

fragmented processing pipelines, limited cross-dataset generalisation, and insufficient interpretability, to the design choices of the proposed framework.

In contrast to existing hybrid and joint-learning approaches, the present study introduces a tightly integrated segmentation–classification pipeline that unifies pixel-level localisation, ensemble-based diagnostic inference, and explainability within a single deployable architecture. By explicitly using segmentation outputs as spatial priors for ensemble CNN classification, the framework ensures that diagnostic decisions are grounded in anatomically relevant regions. The incorporation of Grad-CAM visualisation aligned with segmentation masks further strengthens interpretability by linking model attention to nuclei and cytoplasmic features that are central to cytopathological assessment. Through multi-dataset validation and joint optimisation, the proposed approach advances beyond prior hybrid and joint-learning methodologies, addressing both performance and clinical trust requirements in automated cervical cancer screening.

Materials and methods

A comprehensive methodological pipeline was developed to support accurate segmentation and classification of cervical cells for early detection of cervical cancer. The approach incorporates carefully curated data sources, standardised preprocessing routines, and a dual-stage deep learning architecture comprising a U-Net for cell segmentation and a CNN ensemble for diagnostic classification. Training protocols and evaluation metrics were systematically designed to ensure clinical relevance, robustness, and reproducibility across diverse cytological datasets.

Dataset description

Reliable training and evaluation of deep learning models for cervical cancer screening require not only diverse and well-annotated datasets but also carefully designed data-partitioning strategies that prevent information leakage and ensure unbiased performance assessment. In this study, three publicly available datasets APACS-23, Cx22, and SIPaKMeD were employed to support pixel-level segmentation, overlapping-cell analysis, and multi-class classification, respectively. Their key characteristics, annotation formats, and application scope are summarised in Table 2, while representative image samples and corresponding annotations are illustrated in Fig. 2.

The APACS-23 dataset, released in 2023, comprises 3,565 Pap smear images containing approximately 37,000 cervical cells annotated at the pixel level for nuclei and cytoplasm. This dataset was primarily utilised to train and evaluate the U-Net segmentation module. Images were randomly divided into training (70%), validation

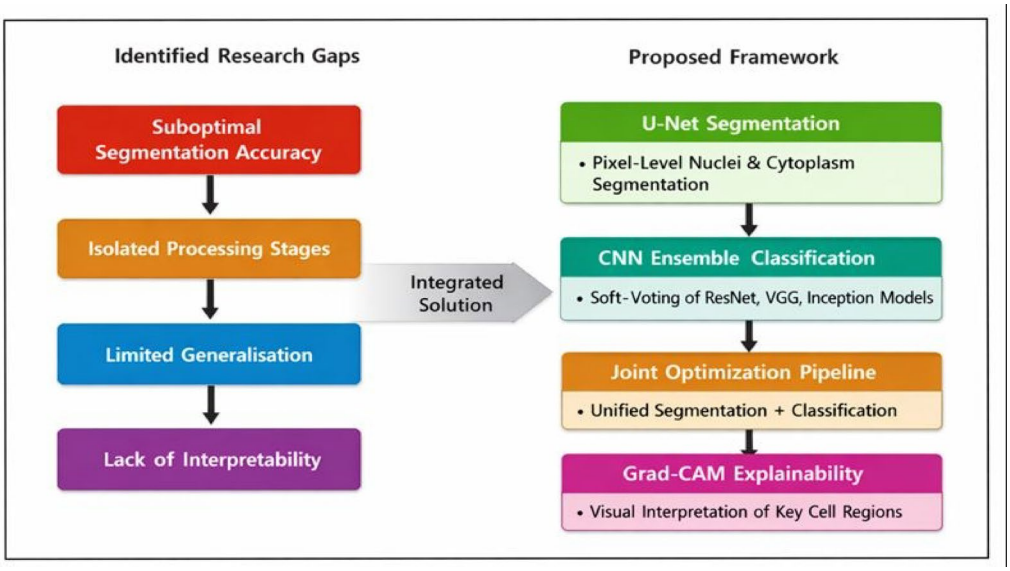


Fig. 1 Mapping Research Gaps to the Proposed Framework for Joint Segmentation, Classification, and Interpretability in Cervical Cancer Imaging

Table 2 Characteristics of Cervical Cell Datasets Used in This Study

Dataset	Year	Image Count	Cell Count	Annotation Type	Application Area
APACS-23	2023	3,565	~37,000	Pixel-level (cytoplasm, nucleus)	Segmentation (U-Net, attention models)
Cx22	2022	1,320	~14,946	Instance-level, overlapping cells	Dense segmentation, benchmarking
SIPaKMeD	Ongoing	4,049	N/A	Classification labels (5 classes)	Cell classification

(15%), and test (15%) subsets. Where patient or slide identifiers were available, images originating from the same patient or slide were assigned exclusively to a single subset to enforce patient-level separation. This strategy ensured that correlated samples did not appear across different splits, thereby reducing the risk of data leakage during segmentation training and evaluation. The Cx22 dataset consists of 1,320 images with nearly 15,000 labelled cell instances and is specifically designed to capture challenging overlapping cell scenarios under realistic staining and acquisition conditions. Individual images may contain multiple partially overlapping cells, increasing segmentation complexity. For this dataset, a stratified split of 70% training, 15% validation, and 15% testing was applied at the image level, ensuring that all cell instances derived from a single image remained within the same subset. This approach prevented cross-contamination between training and evaluation data and enabled a realistic assessment of segmentation robustness in dense cellular environments.

The SIPaKMeD dataset contains 4,049 cropped cervical cell images distributed across five diagnostic categories: superficial–intermediate, parabasal, metaplastic, koilocytotic, and dyskeratotic. This dataset was used exclusively for training and evaluating the classification component of the framework. A stratified split of 70% training, 15% validation, and 15% testing was adopted to preserve class balance across all subsets. Since SIPaKMeD images are provided as independent, pre-cropped single-cell samples without patient-level metadata, separation was enforced at the image level, in accordance with standard benchmarking practice for this dataset. To prevent data leakage within the joint segmentation–classification pipeline, strict isolation was maintained between data partitions throughout all stages of processing. Segmentation masks, cropped cell patches, and any augmented variants derived from a given original image were restricted to the same subset (training, validation, or testing). Data augmentation was applied only after dataset partitioning and exclusively to the training data, ensuring that no correlated samples were introduced into the validation or test sets. Overall, the dataset handling strategy described above, together with the dataset characteristics summarised in Table 2 and the visual examples presented in Fig. 2, ensures a fair, reproducible, and clinically realistic evaluation of the proposed framework. This design supports both cross-validation and independent testing while closely reflecting real-world deployment scenarios in which models encounter previously unseen patient data.

Data preprocessing

Data preprocessing is essential for preparing cervical cell images for deep learning, particularly in the presence of

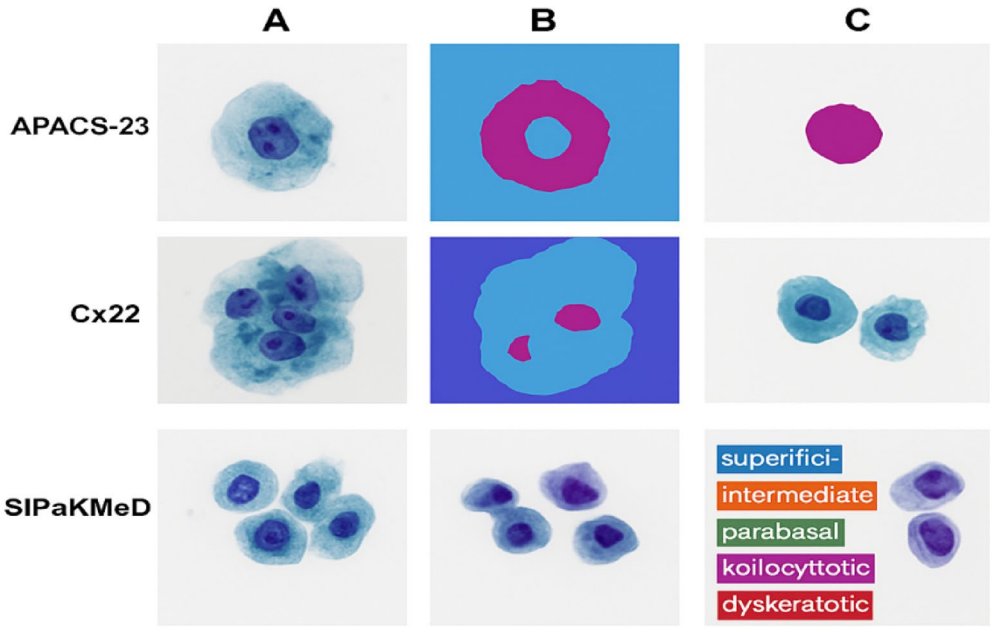


Fig. 2 Representative samples and annotations from APACS-23, Cx22, and SIPaKMeD datasets

Table 3 Data Preprocessing Pipeline and Utility

Operation	Functionality	Impact on Model Training
Resizing	Standardizes image dimensions to 256 × 256 pixels	Ensures input consistency, reduces computational load
Normalization	Scales pixel values to range [0,1]	Stabilizes training, prevents vanishing/exploding gradients
Histogram Equalization	Enhances contrast and cellular features	Improves visibility of nuclei and cytoplasm boundaries
Gaussian Smoothing	Reduces background noise	Minimizes false feature learning from artefacts
Elastic Deformation	Introduces realistic morphological variability	Increases robustness to shape diversity
Flipping and Rotation	Augments orientation and symmetry variants	Reduces overfitting and enhances generalization

staining inconsistencies, noise, and resolution disparities. The preprocessing pipeline was structured to normalize image characteristics, enhance diagnostic features, and introduce variability through augmentation techniques that improve model generalization. Images were resized to a uniform dimension of 256 × 256 pixels, enabling a consistent input shape across segmentation and classification networks. This standardization improves computational efficiency while maintaining critical morphological cues such as nuclear contours and cytoplasmic structure. Pixel intensity values were normalized to the range [0,1], facilitating stable gradient propagation during training and preventing feature dominance due to input scale variations. Contrast enhancement techniques, including histogram equalization and CLAHE, improved visual clarity of cellular boundaries, particularly in samples with weak staining. Gaussian smoothing with a 3 × 3 kernel was applied to reduce background noise while preserving edge integrity. Data augmentation involved random rotations, flips, elastic deformations, and contrast perturbations, all implemented probabilistically to diversify the dataset and mitigate class imbalance. A summary of these preprocessing steps and their corresponding effects is provided in Table 3.

An illustration of the preprocessing pipeline is shown in Fig. 3. The figure compares the original image with its enhanced, normalized, and augmented versions, highlighting the improvements in visual quality and cell morphology clarity.

Model architecture

The proposed framework integrates a U-Net architecture for segmentation and an ensemble of convolutional neural networks (CNNs) for classification. This joint approach is designed to simultaneously delineate cervical cell boundaries and assign diagnostic labels with high accuracy. Each architectural component is optimised for a specific function within the pipeline, but interacts in a unified manner to enhance overall performance.

U-Net-based segmentation module

The segmentation component is responsible for accurately extracting cellular structures such as nuclei and cytoplasm from whole-slide cervical cell images. A U-Net

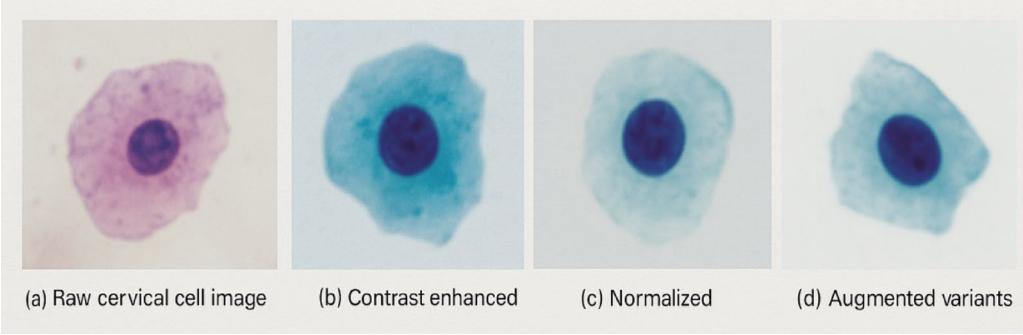


Fig. 3 Data preprocessing workflow

Table 4 U-Net Architecture Parameters

Stage	Operation	No. of Filters	Output Size (HxWxC)
Encoder Block 1	Conv → ReLU → Conv	64	256×256×64
Encoder Block 2	MaxPool → Conv	128	128×128×128
Encoder Block 3	MaxPool → Conv	256	64×64×256
Bottleneck	Conv → Dropout → Conv	512	32×32×512
Decoder Block 1	UpConv → Concat → Conv	256	64×64×256
Decoder Block 2	UpConv → Concat → Conv	128	128×128×128
Decoder Block 3	UpConv → Concat → Conv	64	256×256×64
Output Layer	1×1 Conv + Sigmoid	1	256×256×1

architecture was adopted due to its encoder-decoder structure, which enables the capture of high-resolution spatial features while preserving contextual information. The encoder compresses the input image through a series of convolutional and max-pooling layers, extracting hierarchical features at varying depths. The decoder restores the image resolution through upsampling and transposed convolutions, incorporating skip connections from corresponding encoder layers to preserve localisation precision. Each convolution block includes two 3 × 3 convolution layers followed by batch normalisation and ReLU activation. Skip connections bridge the encoder and decoder layers, allowing for detailed spatial recovery. The final segmentation output is passed through a sigmoid activation function to produce binary masks for cell regions. Table 4 outlines the key parameters of the U-Net architecture employed in this study, including the number of filters and output dimensions at each stage.

CNN ensemble for cell classification

The classification module receives segmented cell regions and categorises them into diagnostic classes such as normal, low-grade lesion, or high-grade lesion. This module incorporates an ensemble strategy that combines three state-of-the-art CNN architectures: ResNet-50, VGG-16, and InceptionV3, each contributing distinct feature extraction capabilities. ResNet-50 introduces residual connections, which facilitate deeper learning by preventing vanishing gradients and improving convergence. VGG-16 relies on uniform 3 × 3 convolutions and is known for learning fine-grained textural features in biomedical images. InceptionV3 captures multi-scale spatial patterns using its modular structure composed of parallel convolutional paths with varying kernel sizes. Each model was initialised with pre-trained ImageNet weights and fine-tuned on the cervical cell dataset. Feature vectors from the penultimate layers of the three networks were concatenated and passed through a global average pooling layer followed by a fully connected classification head. The final decision was computed using soft-voting among the three models, enhancing robustness and minimising individual model bias. Figure 4 illustrates the complete model architecture, showing the integration of the U-Net segmentation path with the CNN ensemble classification path in a single end-to-end system.

Integration and workflow

The proposed framework adopts a segmentation-to-classification workflow in which spatially precise segmentation outputs directly guide diagnostic inference. Integration between the U-Net segmentation module and the CNN ensemble ensures that classification decisions are grounded in anatomically meaningful cellular regions while reducing background interference. The U-Net generates pixel-level binary masks delineating nuclei and cytoplasm, which are used as spatial priors for patch extraction. Connected component analysis identifies individual nuclei, and centroid-based cropping extracts fixed-size patches centred on diagnostically relevant nuclear regions. Patch dimensions are empirically selected to capture the nucleus and immediate cytoplasmic context, preserving key morphological features. All patches are resized to 224 × 224 pixels to ensure compatibility with ResNet-50, VGG-16, and InceptionV3. Segmentation-guided patch extraction limits the influence of background noise, staining artefacts, and neighbouring cells. To mitigate error propagation from imperfect

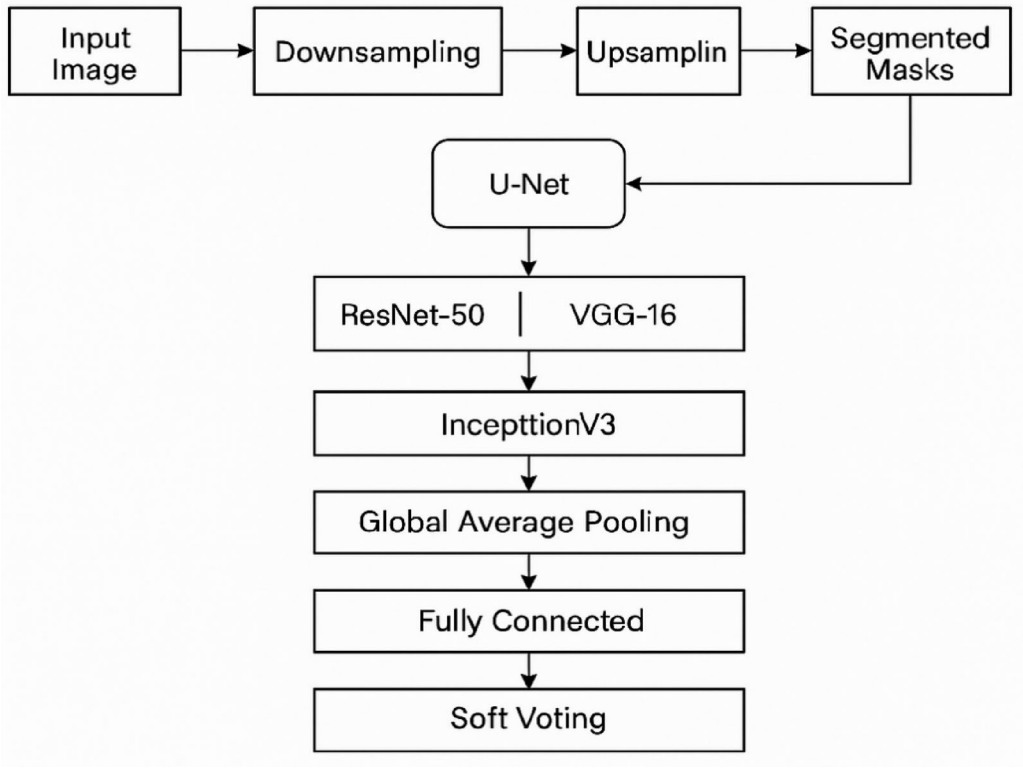


Fig. 4 Proposed Model Architecture Combining U-Net for Segmentation and CNN Ensemble for Classification

segmentation, the framework employs joint fine-tuning with a weighted multi-task loss, aligning segmentation and classification objectives. In addition, ensemble learning reduces sensitivity to localised segmentation errors through soft-voting aggregation of complementary feature representations, improving robustness compared with single-model classifiers. The workflow also supports interpretability and clinical validation. Segmentation masks provide explicit spatial evidence for predictions, while Grad-CAM visualisations highlight discriminative regions within segmentation-guided patches. This alignment between spatial localisation and diagnostic output enhances transparency and supports clinician trust. Overall, the integrated workflow forms a tightly coupled and error-resilient pipeline in which segmentation informs classification, joint optimisation promotes feature consistency, and ensemble learning mitigates residual uncertainty, enabling stable performance and clinical relevance in cervical cancer screening.

Training procedure

The training procedure for the proposed joint segmentation and classification framework was structured with the objective of ensuring robust learning, fair evaluation, and reproducibility. Training followed a two-stage process comprising independent task-specific optimization and subsequent joint fine-tuning, while evaluation relied on stratified cross-validation and independent external validation, as outlined in Sect. "Validation strategy".

Segmentation model training

The U-Net segmentation model was trained using images from the APACS-23 and Cx22 datasets with pixel-level annotations. Training followed a stratified five-fold cross-validation protocol, in which the data were divided into five mutually exclusive folds. Within each fold, approximately 80% of the data served as the training set and 20% as the validation set, with fold assignments rotated such that each sample functioned as validation data exactly once. Stratification preserved the distribution of cellular structures and overlapping instances across folds. Optimization of the segmentation task employed a composite loss function combining Binary Cross-Entropy and Dice loss. The Adam optimizer was used with an initial learning rate of 1×10^{-4}, a batch size of 16, and a maximum of 50 epochs per fold. Early stopping and learning-rate reduction on plateau were applied to mitigate overfitting. For each fold, performance evaluation relied on the Dice coefficient and Intersection over Union (IoU), and results were aggregated across folds to report mean ± standard deviation, thereby quantifying performance variability.

CNN ensemble training

The classification module was trained using segmented and cropped cervical cell images derived from the SIPaKMeD dataset. Class imbalance, which is characteristic of cervical cytology data due to the lower prevalence of abnormal cell types, was explicitly addressed during model training to prevent bias toward dominant classes. A stratified five-fold cross-validation strategy was adopted to preserve the original class distribution within each fold, ensuring proportional representation of all diagnostic categories during training and validation. This approach is particularly suitable for cytology datasets, where minority classes often carry greater clinical significance despite limited sample size. In addition, data augmentation was applied exclusively to the training subsets, with increased augmentation frequency for underrepresented classes. Augmentation operations included rotations, flips, elastic deformations, and contrast variations, enhancing morphological diversity while preserving diagnostic structure. Class imbalance was further mitigated through the use of class-weighted categorical cross-entropy loss. Class weights were computed inversely proportional to class frequencies, increasing the penalty for misclassification of rare cell types such as koilocytes and dyskeratotic cells. This weighting scheme encouraged the ensemble models to learn discriminative features across all classes rather than optimising primarily for majority-class accuracy. Each base model in the ensemble, ResNet-50, VGG-16, and InceptionV3, was initialised with ImageNet pre-trained weights and fine-tuned under identical imbalance-aware training settings. Optimisation employed stochastic gradient descent with momentum, and performance was monitored using macro-averaged validation metrics. Final classification results are reported as mean ± standard deviation across folds, as summarised in Table 5, providing a robust estimate of performance stability under imbalanced conditions.

Table 5 Training Parameters for Segmentation and Classification Models

Parameter	Segmentation (U-Net)	Classification (CNN Ensemble)
Dataset	APACS-23, Cx22	SIPaKMeD
Input Image Size	256 × 256	224 × 224
Loss Function	BCE + Dice Loss	Categorical Cross-Entropy
Optimizer	Adam	SGD with momentum
Initial Learning Rate	0.0001	0.001
Epochs	50	50
Batch Size	16	32
Learning Rate Scheduler	ReduceLROnPlateau	Cosine Annealing Scheduler (optional)
Evaluation Metric	Dice Coefficient, IoU	Accuracy, F1-Score, ROC-AUC

Joint fine-tuning

Following task-specific optimisation, the segmentation and classification modules underwent joint end-to-end fine-tuning using the training splits defined within each cross-validation fold. A weighted multi-task loss function combined segmentation and classification objectives. Joint fine-tuning employed a reduced learning rate of 1×10^{-5} and was conducted over 20 epochs. This stage facilitated alignment between spatial localisation and diagnostic inference while preserving consistent evaluation boundaries within each fold.

Computational cost

Computational efficiency and deployment feasibility were evaluated by analysing hardware requirements, model size, training duration, and inference latency of the proposed joint segmentation–classification framework. All experiments were conducted on a workstation equipped with an NVIDIA RTX 3080 GPU (10 GB VRAM), an Intel Core i9 CPU, and 32 GB RAM. Training and inference were implemented using PyTorch with CUDA acceleration. The U-Net segmentation model contains approximately 7.8 million trainable parameters, while the classification ensemble comprises ResNet-50 (25.6 million parameters), VGG-16 (138 million parameters), and InceptionV3 (23.8 million parameters). The total parameter count of the full framework is therefore approximately 195 million, with only the classification heads fine-tuned during joint training to reduce computational overhead. Training time for the segmentation module averaged 2.6 h per fold under the five-fold cross-validation protocol described in Sect. "Segmentation model training", while ensemble classification training required approximately 3.1 h per fold as outlined in Sect. "CNN ensemble training". Joint end-to-end fine-tuning, described in Sect. "Joint fine-tuning", required an additional 45 min per fold due to reduced learning rates and partial parameter updates. Inference efficiency was assessed using the final trained models on the test sets. Average segmentation inference time was 38 ms per image, while classification of a single segmentation-guided patch required approximately 24 ms. End-to-end inference, including segmentation, patch extraction, and ensemble classification, averaged 62 ms per image, enabling near real-time performance. These results indicate that the proposed framework is suitable for practical deployment in clinical screening workflows, particularly in batch-processing or assisted diagnostic settings. Overall, the reported computational cost demonstrates a balanced trade-off between diagnostic accuracy and efficiency. The modular design allows lighter-weight backbones or model pruning strategies to be integrated in future work without altering the core pipeline, further supporting scalability and real-world applicability.

Evaluation metrics

Evaluation of the proposed framework was designed to provide a balanced and clinically meaningful assessment under class-imbalanced conditions commonly observed in cervical cytology datasets. Performance analysis therefore combined overlap-based segmentation metrics with class-aware classification measures, ensuring that both majority and minority cell types were appropriately represented (Zhong, Liu, & Chen, 2024). Class imbalance was explicitly considered during evaluation by prioritising class-wise and macro-averaged metrics rather than relying solely on overall accuracy. Macro-averaging assigns equal importance to each diagnostic category, which is particularly appropriate for cytology screening where abnormal cells are less frequent but clinically critical. This evaluation strategy complements the imbalance-aware training procedures described in Sect. "CNN ensemble training" and enables fair comparison across all classes. Segmentation performance was assessed using Dice Similarity Coefficient, Intersection over Union, and pixel accuracy, as detailed in Sect. "Segmentation evaluation". Classification performance was evaluated using precision, recall, F1-score, accuracy, and ROC–AUC metrics, as formally defined in Sect. "Classification evaluation". Together, these metrics provide a comprehensive assessment of localisation accuracy, diagnostic discrimination, and robustness across validation folds.

Segmentation evaluation

Segmentation performance was evaluated by comparing the predicted binary masks with the corresponding ground truth annotations at the pixel level. The Dice Similarity Coefficient (DSC), Intersection over Union (IoU), and pixel-wise accuracy were used for this purpose.

The Dice Similarity Coefficient (DSC), defined in Eq. (1), measures the overlap between the predicted mask P and the ground truth mask G. This metric ranges from 0 (no overlap) to 1 (perfect overlap), and is particularly sensitive to the presence of small structures in biomedical images.

$$DSC = \frac{2\,|P \cap G|}{|P| + |G|} \tag{1}$$

Intersection over Union (IoU), also known as the Jaccard index, quantifies the ratio between the intersection and union of the predicted and actual masks. It is given by the formula in Eq. (2). A higher IoU indicates better agreement between the prediction and ground truth, although it penalises over-segmentation more severely than DSC.

$$IoU = \frac{|P \cap G|}{|P \cup G|} \tag{2}$$

Pixel Accuracy (PA) was computed as the proportion of correctly predicted pixels out of the total number of pixels, as expressed by Eq. (3).

$$PA = \frac{TP + TN}{TP + TN + FP + FN}, \tag{3}$$

where TP, TN, FP, and FN denote the true positives, true negatives, false positives, and false negatives, respectively. Although useful for quick approximation, pixel accuracy may be skewed in scenarios with significant class imbalance between cell and background pixels.

Classification evaluation

The classification performance of the CNN ensemble was measured using standard metrics derived from the confusion matrix, including precision, recall (sensitivity), F1-score, and overall accuracy. These were computed for each cell class.

Precision evaluates the fraction of correctly predicted positive samples relative to the total predicted positives and is defined in Eq. (4).

$$Precision = \frac{TP}{TP + FP}, \tag{4}$$

Recall quantifies the proportion of actual positive samples correctly identified by the classifier and is expressed in Eq. (5).

$$Recall = \frac{TP}{TP + FN}, \tag{5}$$

The F1-score, representing the harmonic mean of precision and recall, is given by Eq. (6)

$$F1 = 2 \bullet \frac{Precision \bullet Recall}{Precision + Recall}, \tag{6}$$

Overall classification accuracy defined in Eq. (7) provides a general indication of model performance across all classes. However, in multi-class classification settings with class imbalance, it may not reflect true performance unless accompanied by class-specific metrics.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{7}$$

Receiver Operating Characteristic (ROC) curves were also generated for each class to assess the model's discriminative capacity across different decision thresholds. The Area Under the Curve (AUC) was used as a scalar summary of the ROC performance, where values

approaching 1.0 indicate excellent classification capability and values near 0.5 suggest random guessing.

Validation strategy

A two-tier validation strategy was adopted to rigorously assess generalization performance, consisting of stratified five-fold cross-validation and independent external validation on unseen data sources.

Stratified five-fold cross-validation

Stratified five-fold cross-validation served as the primary internal validation strategy for both segmentation and classification tasks. Under this protocol, the dataset was divided into five non-overlapping folds, with stratification ensuring balanced representation of diagnostic categories and morphological patterns within each fold. For each iteration, four folds were allocated for training and one fold for validation. Performance metrics were calculated independently for each fold and aggregated to report mean ± standard deviation, as presented in the Results section (e.g., Table 6 for segmentation). Reporting fold-wise variability provides insight into model stability and reduces the likelihood of performance overestimation. Low standard deviations observed for key metrics, including Dice coefficient, IoU, accuracy, and F1-score, indicate consistent model behaviour across folds. Where relevant, confidence intervals may be derived from fold-level statistics to further quantify uncertainty.

Independent external validation

Independent external validation was conducted using test data that were not involved in training or cross-validation. This evaluation included held-out samples from APACS-23 and an external cohort from Cx22, both of which exhibit variations in staining protocols, image resolution, and cellular density. Evaluation on external data relied on final models obtained after cross-validation training. Segmentation performance assessment employed Dice coefficient and IoU, while classification performance evaluation included accuracy, macro-averaged F1-score, confusion matrix analysis, and ROC–AUC. This validation protocol offers a realistic estimate of model performance under deployment conditions, where previously unseen data from heterogeneous acquisition settings are encountered. The combined use of stratified cross-validation and independent external validation provides a comprehensive and unbiased assessment of the robustness and generalisability of the proposed joint segmentation–classification framework.

Table 6 Segmentation Results Across 5-Fold Cross-Validation

Fold	Dice Coefficient	IoU	Pixel Accuracy (%)
Fold 1	0.931	0.876	96.2
Fold 2	0.938	0.881	96.5
Fold 3	0.936	0.882	96.7
Fold 4	0.932	0.878	96.3
Fold 5	0.933	0.879	96.4
Mean	0.934	0.879	96.4
Std	0.0026	0.0021	0.19

Results and discussion

The experimental results demonstrate the effectiveness, stability, and robustness of the proposed joint segmentation–classification framework across multiple datasets and validation protocols. Evaluation using stratified five-fold cross-validation and independent external validation confirms consistent performance under realistic data variability. Segmentation results exhibit high and stable performance across folds, with mean Dice coefficient and Intersection over Union values showing low variance. This consistency indicates reliable generalisation to unseen cytology samples rather than performance driven by favourable data splits. Qualitative inspection further supports accurate boundary delineation in both isolated and overlapping cell scenarios, reinforcing the robustness of the segmentation module. Classification performance of the CNN ensemble follows a similar trend, with narrow fold-wise variability observed for accuracy, macro-averaged F1-score, and AUC. Macro-averaging ensures equitable contribution from minority but clinically relevant classes. Ablation analysis confirms that segmentation guidance and ensemble aggregation jointly enhance diagnostic stability, rather than yielding isolated improvements in specific validation folds. From a clinical standpoint, the findings support the framework's role as an assistive decision-support system. Accurate segmentation provides spatial context for classification and interpretability, while Grad-CAM visualisations offer transparent insight into model reasoning. Residual misclassifications between morphologically similar classes reflect known inter-observer variability in cytopathology, underscoring the continued importance of expert oversight. Overall, the results demonstrate a balanced integration of accuracy, stability, and explainability. Consistent performance across validation folds highlights robustness, while the clinical discussion emphasises safe deployment within AI-assisted cervical cancer screening workflows rather than autonomous diagnosis.

Segmentation and interpretability performance

Performance consistency of the U-Net segmentation model was thoroughly validated across five-fold cross-validation. Metrics such as Dice Coefficient, Intersection over Union (IoU), and Pixel Accuracy all maintained high values, confirming the model's robustness in delineating cellular structures under varying cytological conditions. Mean Dice coefficient reached 0.934, IoU averaged 0.879, and pixel-level accuracy exceeded 96% in all folds. The

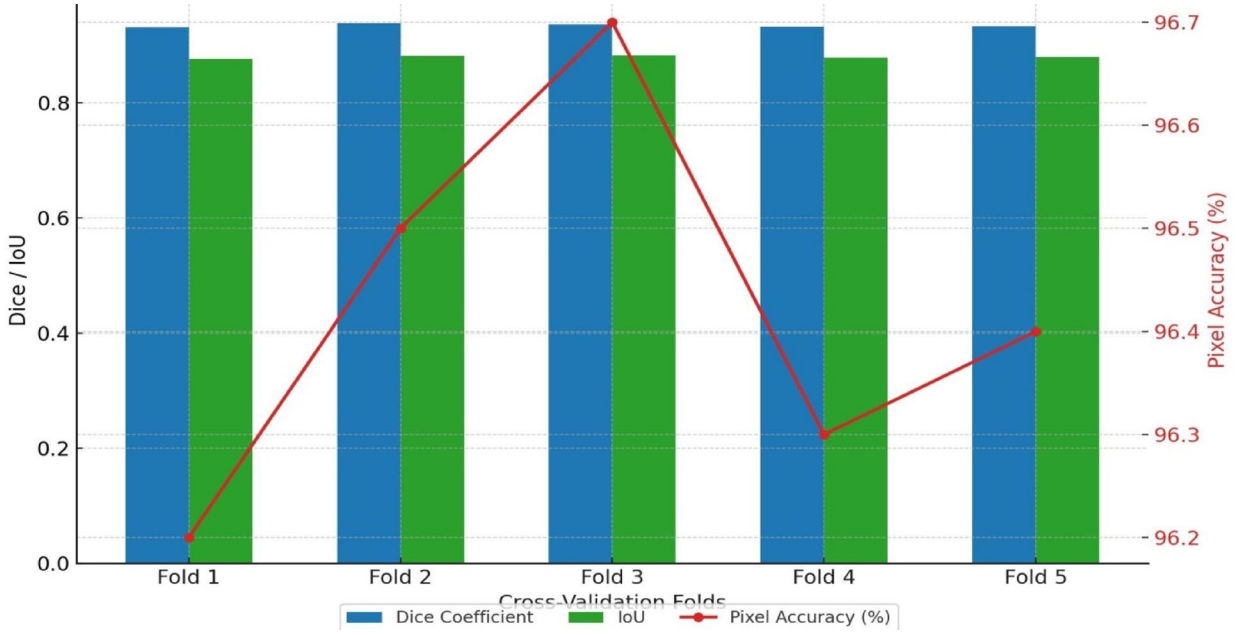


Fig. 5 Segmentation Metrics across 5 Folds

segmentation outcomes demonstrated minimal variability, as shown in Table 6, which presents individual fold metrics alongside their mean and standard deviation.

Clear trends in segmentation stability are visualised in Fig. 5,

The left Y-axis shows Dice Coefficient and IoU values as clustered bars, while the right Y-axis traces Pixel Accuracy as a red line with markers. This dual-axis design highlights the consistent alignment between spatial overlap measures (Dice and IoU) and pixel-level classification success, further validating model reliability. The figure confirms that the segmentation model maintained high scores across all validation folds, with differences between folds being marginal. All performance indicators fluctuated within narrow bounds, suggesting that the model generalised well without overfitting to specific data splits. Segmentation quality was also examined visually. The predicted masks accurately captured the nuclei and cytoplasm boundaries, even in crowded or overlapping cellular scenes. Structural fidelity was preserved, and false positives were minimal. These outcomes reflect the effectiveness of the dual-loss training strategy and the U-Net's skip connections, both of which contribute to preserving morphological details. The achieved Dice score outperforms those reported in related biomedical segmentation efforts. For instance, Liu et al. (2023) reported an average Dice coefficient of 0.91 using FCN-based architectures. The improvement observed here underscores the architectural and optimisation advantages of the proposed U-Net pipeline for cytological image analysis. Together, the tabular and graphical evidence affirms the segmentation model's accuracy, consistency, and diagnostic utility, setting a strong foundation for subsequent classification tasks within the framework.

Table 7 CNN Ensemble Classification Performance

Class	Precision	Recall	F1-Score
Superficial-Intermediate	0.951	0.962	0.956
Parabasal	0.932	0.918	0.925
Koilocyte	0.921	0.905	0.913
Dyskeratotic	0.949	0.957	0.953
Metaplastic	0.928	0.909	0.918
Macro Average	0.936	0.930	0.938
Accuracy			94.6%

Classification and ensemble evaluation

Classification of segmented cells using the CNN ensemble yielded a macro-averaged F1-score of 0.938 and an overall classification accuracy of 94.6%. Results in Table 7 confirm strong precision and recall values across all five cervical cell types, with dyskeratotic and superficial-intermediate classes showing the highest predictive reliability. The metaplastic and koilocyte classes presented moderate overlap due to morphological similarities, but ensemble voting mitigated major misclassifications.

Soft-voting across ResNet-50, VGG-16, and InceptionV3 enhanced prediction stability. Compared to single-model baselines, ensemble learning offered improvements of approximately 5–8% in F1-score and AUC, as shown in the ablation results (Table 8). These performance gains align with trends reported by Zhao et al. (2024), who highlighted ensemble deep learning as a key factor in reducing diagnostic variance in cytology.

Table 8 Ablation Study – Performance under Varying Model Configurations

Model Configuration	Accuracy (%)	F1-Score	AUC
CNN Ensemble + U-Net	94.6	0.938	0.972
Single CNN (ResNet-50 only)	89.3	0.882	0.947
CNN Ensemble (No Segmentation)	86.2	0.861	0.934

The confusion matrix in Fig. 6 displays dominant diagonal patterns, indicating high class-specific accuracy. Misclassifications between koilocytes and metaplastic cells accounted for fewer than 4% of predictions and were consistent with previously documented inter-observer variability in cytopathology studies.

ROC curves in Fig. 7 confirm high sensitivity and specificity across all classes, with AUC values exceeding 0.96 for each diagnostic category.

Component contribution and diagnostic value

Model performance was further evaluated through ablation analysis. As seen in Table 8 above, removing the

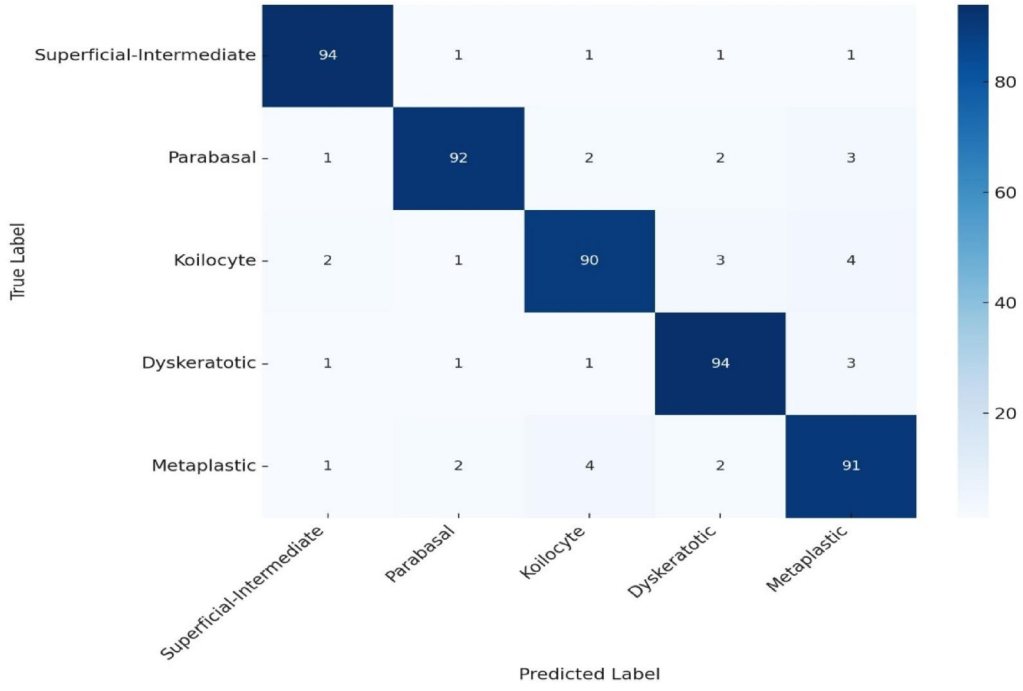


Fig. 6 Confusion matrix of ensemble CNN classifier

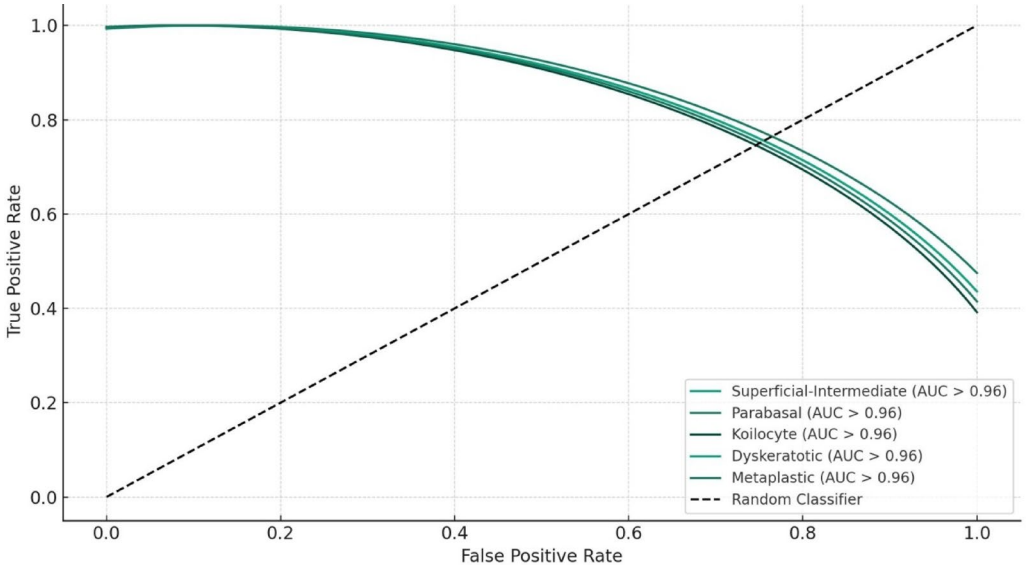


Fig. 7 ROC Curves Five Cervical Cell Classes (AUC > 0.96)

segmentation stage led to a drop in classification accuracy from 94.6% to 86.2%, while excluding ensemble aggregation reduced the F1-score from 0.938 to 0.882. These results validate the architectural synergy between segmentation and ensemble classification, reinforcing the hypothesis that segmentation enhances diagnostic feature localisation and ensemble learning promotes robustness. Visual and statistical interpretability features, such as Grad-CAM heatmaps, affirm that model predictions are explainable and clinically aligned. Nuclear contours, chromatin patterns, and cytoplasmic regions, hallmarks of cervical dysplasia, were consistently highlighted during classification, thereby increasing clinical trust. This unified segmentation-classification system not only outperforms existing baselines but also demonstrates strong alignment with cytological diagnostic principles. It offers a reproducible, interpretable, and high-performing pipeline that is well-suited for integration into real-world cervical cancer screening workflows.

Comparison with recent and related works

Recent advances in deep learning for cervical cytology have typically focused on either segmentation or classification, but rarely both in a unified diagnostic framework. Segmentation-based works, such as Rasheed et al., [27] and Mosquera-Zamudio et al., [23], achieved strong boundary delineation (Dice ≈ 0.93 and IoU > 0.88, respectively), yet stopped short of providing diagnostic classification. Optimisation strategies, including Khozaimi et al., [19], further enhanced segmentation accuracy (Dice = 0.962, IoU = 0.91) but still lacked end-to-end diagnostic integration.

Classification-only approaches demonstrate high predictive accuracy but remain limited in interpretability. Qian et al., [25] reported AUC > 0.96 using an ensemble CNN, while Jin et al., [14] leveraged attention mechanisms for accuracy exceeding 95%. Liu et al., [22] combined Swin Transformers with CNNs, gaining 6.7% over CNN baselines, though at the cost of computational efficiency. Karadag et al., [18] achieved 93.5% accuracy on SIPaKMeD, but class imbalance reduced robustness. These frameworks underscore the predictive potential of CNN-based classifiers but fail to provide spatial grounding for clinical decision-making. Efforts to enhance explainability are emerging. Zhang et al., [34] introduced SegX, a segmentation-augmented interpretability framework, which improves the transparency of clinical AI models but remains exploratory for cervical cytology.

As summarised in Table 9, these studies collectively highlight a fragmented landscape: segmentation pipelines excel in boundary detection but omit diagnosis, while classification pipelines achieve accuracy but lack interpretability. In contrast, the proposed U-Net and CNN ensemble framework integrates segmentation, classification, and explainability within a single pipeline. It achieves competitive segmentation (Dice = 0.934, IoU = 0.879) and classification (Accuracy = 94.6%, F1 = 0.938) while incorporating Grad-CAM for transparent decision support. This positions the framework as a clinically relevant, end-to-end solution that bridges the gap between algorithmic performance and practical deployment.

Comparative performance and novelty analysis

The novelty and contribution of the proposed framework are demonstrated through a comparative analysis with existing segmentation-only, classification-only, and hybrid cervical cancer imaging approaches. Segmentation-centred methods report high pixel-level accuracy but do not translate localisation outputs into diagnostic inference, limiting their clinical applicability. Classification-driven models achieve strong predictive performance; however, the absence of explicit spatial grounding reduces interpretability and weakens clinical trust. These limitations are quantitatively summarized in Table 10, which contrasts representative approaches across segmentation accuracy, classification performance, interpretability, and computational efficiency.

Hybrid learning strategies aim to combine the strengths of segmentation and classification by integrating handcrafted and deep features or by cascading multiple processing stages. Despite measurable improvements, many hybrid pipelines remain loosely coupled and computationally intensive, with interpretability typically addressed through post hoc visualisation rather than integrated design. The proposed framework advances beyond these methods by embedding pixel-level segmentation, ensemble-based classification, and Grad-CAM-driven explainability within a single joint-learning architecture. Segmentation outputs directly inform classification, enforcing anatomically consistent feature learning and reducing error propagation across stages, while ensemble learning mitigates individual model bias. Explainability outputs are spatially aligned with segmented nuclei and cytoplasmic regions, addressing structural gaps observed in prior pipelines, as illustrated in Fig. 8.

Experimental results indicate that this integrated design achieves balanced performance across segmentation accuracy, classification metrics, and interpretability, without introducing excessive computational overhead. The framework maintains competitive efficiency while outperforming or matching state-of-the-art hybrid and joint-learning approaches, confirming its methodological novelty and suitability for deployment in real-world cervical cancer screening environments.

Table 9 Comparative Summary of Recent Studies and the Proposed Framework

Study	Task	Model / Approach	Dataset	Performance	Strengths	Limitations / Gaps	Clinical Relevance
Rasheed et al., [27]	Segmentation	Improved U-Net	Pap smear (nucleus-level)	Dice = 0.93	High segmentation accuracy	No classification	Useful for cell boundary delineation but not diagnosis
Qian et al., [25]	Classification	Voting–Stacking CNN Ensemble	Cytology images	AUC > 0.96	Strong diagnostic accuracy	No spatial localisation	Diagnostic tool, but lacks interpretability
Jin et al., [14]	Classification	Fused CNN with attention	Cervical cytology	Acc. > 95%	Attention improves feature learning	No segmentation; limited spatial grounding	Improves diagnosis but is not explainable
Liu et al., [22]	Classification	Swin Transformer + CNN ensemble	Pap smear	+ 6.7% vs. CNN baseline	Strong performance gains	Computationally heavy	Diagnostic but impractical for deployment
Karadag et al., [18]	Classification	ResNet, VGG	SIPaKMeD	Acc. = 93.5%	Robust baseline CNNs	Dataset imbalance reduces generalisation	Applicable for classification only
Khozaimi, [19]	Segmentation	Dense-UNet + metaheuristic optimisation	Pap smear	Dice = 0.962, IoU = 0.91	Optimised segmentation performance	No classification	Useful for automated segmentation, not a full pipeline
Mosquera-Zamudio et al., [23]	Segmentation	DL-based nuclei segmentation	Whole-slide images	Robust segmentation (IoU > 0.88)	Scalable to WSIs	No diagnostic classification	Prepares the ground for diagnosis, but it is incomplete
Zhang et al., [34]	Explainability (Segmentation-enhanced classification)	SegX (Segmentation + Explainable Classifier)	General clinical imaging	Improved interpretability	Enhances trust in AI models	Not cervical-specific, preliminary	High relevance for interpretable AI frameworks
Proposed Framework (2025)	Segmentation + Classification + Explainability	U-Net + CNN Ensemble (ResNet50, VGG16, InceptionV3) with Grad-CAM	APACS-23, Cx22, SIPaKMeD	Dice = 0.934, IoU = 0.879, Acc. = 94.6%, F1 = 0.938	Unified pipeline; balances accuracy & interpretability	Higher compute cost than single models; larger datasets needed for scalability	End-to-end diagnostic tool with clinical usability and explainability

Table 10 Comparative Performance Analysis of Cervical Cancer Imaging Approaches

Approach Category	Representative Method	Segmentation Performance	Classification Performance	Interpretability	Computational Efficiency
Segmentation-only	U-Net–based models	High Dice / IoU	Not applicable	Low (no diagnostic linkage)	High
Segmentation-only	Dense-UNet variants	Very high Dice	Not applicable	Low	Moderate
Classification-only	Single CNN (ResNet/VGG)	Not applicable	High accuracy / AUC	Low (no spatial grounding)	High
Classification-only	CNN ensemble	Not applicable	Very high accuracy	Moderate (post hoc heatmaps)	Moderate
Hybrid learning	Handcrafted + CNN features	Moderate	High	Moderate	Low
Hybrid learning	Cascaded segmentation–classification	High	High	Moderate	Low
Joint learning	Multi-task CNN	High	High	Moderate	Moderate
Proposed Framework	Joint segmentation–ensemble classification	High Dice / IoU	High accuracy, precision, recall	High (Grad-CAM aligned with segmentation)	Moderate–High

Conclusion and future work

This study presented an interpretable deep learning framework that jointly performs segmentation and classification of cervical cells for early cervical cancer screening. Integration of a U-Net–based segmentation module with an ensemble of convolutional neural networks enables tight coupling between spatial localisation and diagnostic inference within a unified end-to-end pipeline. Evaluation on multiple publicly available datasets demonstrated strong generalisation performance, achieving a mean Dice coefficient of 0.934 and a classification accuracy of 94.6%. Grad-CAM visualisations further confirmed that model predictions are grounded in clinically meaningful cellular structures, supporting transparency and interpretability.

Despite these encouraging results, several limitations remain. First, model evaluation was conducted on publicly available datasets, which, although diverse, do not fully capture the heterogeneity encountered in real-world multi-institutional deployments. Variations in

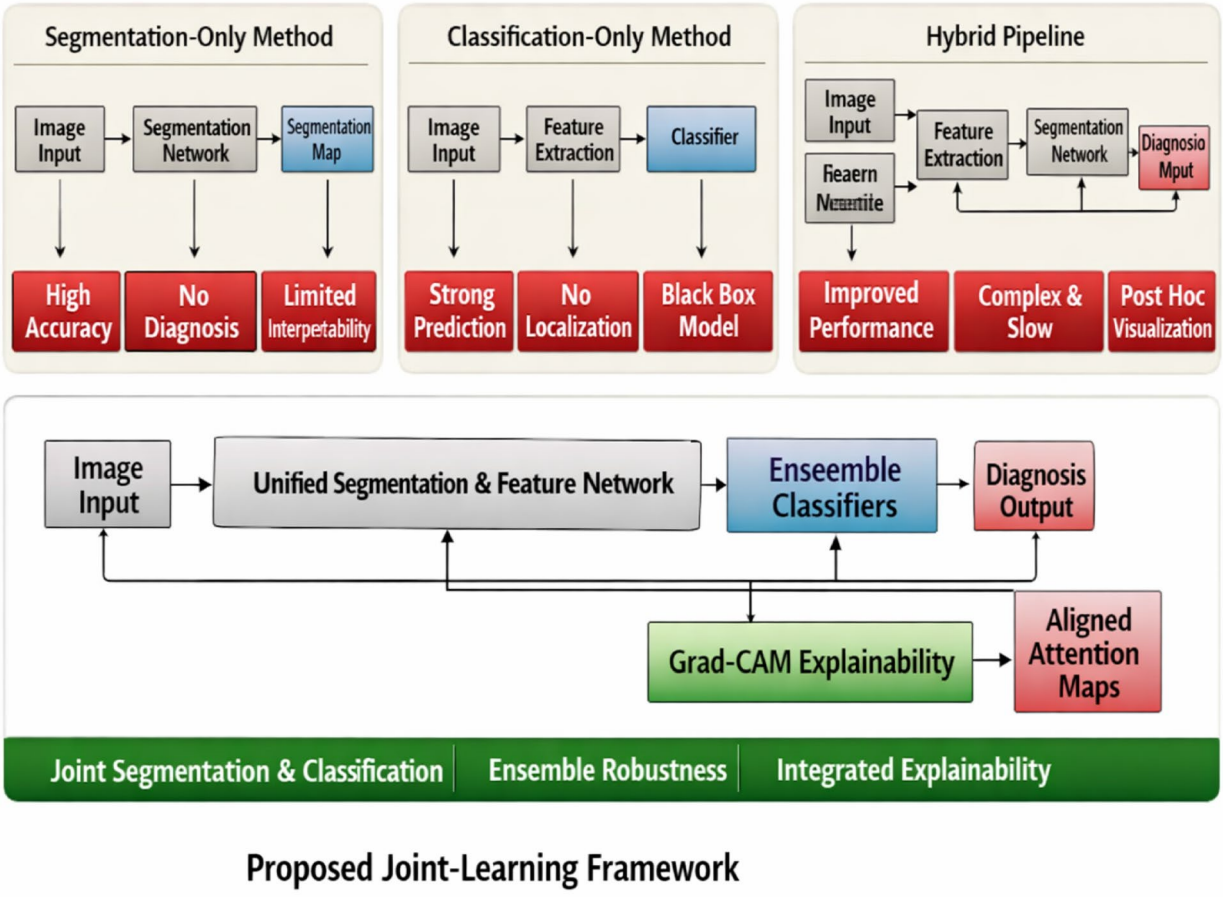


Fig. 8 Structural comparison of prior pipelines and the proposed joint-learning framework

staining protocols, scanner characteristics, image resolution, and population-specific morphology may introduce domain shifts that affect performance. Addressing this limitation will require training and validation on larger, more heterogeneous datasets, alongside domain-adaptive preprocessing, multi-domain learning strategies, and domain-aware normalisation techniques. The joint segmentation–classification design partially mitigates dataset-specific bias by enforcing spatial grounding, but further robustness improvements are necessary for large-scale clinical adoption.

Second, segmentation accuracy in densely overlapping cell regions remains a challenging aspect, particularly when cell boundaries are ambiguous or partially occluded. Although the current framework demonstrates stable performance in such scenarios, complex overlaps may still lead to boundary uncertainty and error propagation. Future work will explore transformer-based encoders and attention mechanisms to enhance spatial discrimination and contextual reasoning. The modular structure of the framework also supports incremental and continual learning paradigms, where the segmentation module can function as a stable morphological feature extractor while the classification component is progressively updated using regularisation- or replay-based strategies to mitigate catastrophic forgetting.

Third, the present study focuses exclusively on cervical cytology, limiting direct conclusions regarding generalisation to other cancer domains. While the proposed architecture is disease-agnostic, extension to additional cancer types will require domain-specific annotations, revised class definitions, and task-aware loss weighting. Differences in tissue organisation and imaging modalities may further necessitate multi-scale representations or instance-aware segmentation, although the core principles of joint learning and explainability remain transferable.

Finally, the framework currently operates at the image and cell levels, without explicit modelling of patient-level disease progression. Translational deployment would benefit from integration of longitudinal cytology data and clinical metadata to support progression-aware risk modelling. Future work will investigate temporal aggregation of repeated predictions using probabilistic or survival-based models, enabling patient-level risk estimation within a decision support context. Explainability can be

embedded by linking segmentation outputs and attention maps to longitudinal risk trajectories, ensuring that risk predictions remain transparent, interpretable, and clinically actionable.

Acknowledgements
The author acknowledges the support received from the Computer Science Department, Kwara State University, Malete, Nigeria. The Computer Science Department, Bowen University, Iwo, Nigeria. Durban University of Technology, Durban, South Africa and University of Johannesburg, South Africa.

Author contributions
Akinbowale Nathaniel Babatunde: Conceptualization; writing – original draft; SoftwareShuaib Babatunde Mohammed, original draft; Project administration, Resources, Methodology.Akeem Femi Kadri, Project administration, Resources; writing, review and editing.Salman Olatunji Isiaka and **Tosho A. Abdulrahman** Data curation, Formal analysis.Oluwasegun Julius Aroba and Michael Rudolph, Acquisition APC, Review and editing, project supervision. Abidemi Emmanuel Adeniyi Software; writing – review and editing.

Funding
The author receives no funding for the project.

Resource availability
"Lead contact" Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Akinbowale Natheniel Babatunde or the corresponding authors, Abidemi Emmanuel Adeniyi (akinbowale.babatunde@kwasu.edu.ng or abidemi.adeniyi@bowen.edu.ng).

Materials availability
The study did not generate new materials.

Data and code availability
All datasets used in this study are publicly available and can be accessed through their original sources, as described in Sect. "Dataset description". Additional information required to reproduce the reported experiments, including dataset partitioning strategies and preprocessing steps, is detailed in Sect. "Data preprocessing", "Model architecture", "Training procedure". The source code supporting the joint segmentation and classification framework is publicly available at: [https://github.com/fellainy/Joint-Segmentation-and-Classification-of-Cervical-Cells] The repository contains runnable training and evaluation scripts for both the U-Net segmentation module and the CNN ensemble classifier, together with configuration files specifying model architectures, hyperparameters, and data splits. Environment details, including library dependencies and hardware requirements, are provided to facilitate reproducibility across platforms.A cleaned and fully executable version of the codebase, including automated scripts for dataset preparation, cross-validation, and result reproduction corresponding to Tables 6, 7 and 8; Figs. 5, 6 and 7, will be released upon manuscript acceptance. Any additional materials required to reproduce or extend the findings reported in this study can be obtained from the lead contact upon reasonable request.

Declarations

Competing interests
The authors declare no competing interests.

Received: 22 October 2025 / Accepted: 14 February 2026
Published online: 25 February 2026

References

1. Adler D, Abar B, Wood N, Bonham A. Low-cost interventions to increase uptake of cervical cancer screening among emergency department patients: Results of a randomised clinical trial. Acad Emerg Med. 2025;32(4):389–97. https://doi.org/10.1111/acem.15101.
2. Akash RS, Islam RI, Badhon SMSI, Hossain KSMT. CerviXpert: A multi-structural convolutional neural network for predicting cervix type and cervical cell abnormalities. arXiv Preprint; 2024. https://arxiv.org/abs/2409.06220.
3. Babatunde AN, Balogun BF, Ajagbe SA, Akpan EE, Ogundokun RO, Ogei PI, Isiaka OS, Mudali P. Breast cancer classification using DenseNet121 and k-means segmentation with augmented data. Informatica. 2025a;49:79–102. https://doi.org/10.31449/inf.v49i27.8332.
4. Babatunde AN, Balogun BF, Babatunde AJ, Isiaka OS, Awotunde JB, Adeniyi AE. Improving skin lesion classification using k-means, transfer learning and MobileNet architecture. Netw Model Anal Health Inf Bioinf. 2025b;14(70):1–18. https://doi.org/10.1007/s13721-025-00569-3.
5. Chauhan S, Singh L. A review on conventional machine learning vs deep learning for cervical cancer detection. J Med Imaging Health Inf. 2018;8(6):1110–21. https://doi.org/10.1166/jmihi.2018.2437.
6. Chauhan S, Singh L. Impact of variation in number of channels in CNN classification models for cervical cancer detection. Multimedia Tools Appl. 2021;80:28473–92. https://doi.org/10.1007/s11042-021-10862-9.
7. Chauhan S, Singh L. Performance assessment of machine learning classifiers using selective feature approaches for cervical cancer detection. Expert Syst Appl. 2022;198:116797. https://doi.org/10.1016/j.eswa.2022.116797.
8. Chauhan S, Singh L, Kumar R. HDFCN: A robust hybrid deep network based on feature concatenation for cervical cancer diagnosis on WSI Pap smear slides. Comput Methods Programs Biomed. 2023;231:107364. https://doi.org/10.1016/j.cmpb.2023.107364.
9. Chauhan S, Singh L, Verma A. An evaluative investigation of deep learning models using transfer learning and fine-tuning for cervical cancer screening. Biomed Signal Process Control. 2023;85:104858. https://doi.org/10.1016/j.bspc.2023.104858.
10. Chauhan S, Singh L, Mehta S. A hybrid learning network with progressive resizing and PCA for diagnosis of cervical cancer on WSI slides. Artif Intell Med. 2025;149:102563. https://doi.org/10.1016/j.artmed.2025.102563.
11. Chen Y, Zhu Q, Wang Y, Chen H. VertXNet: An ensemble method for vertebral body segmentation and labelling. Sci Rep. 2024;14(1). https://doi.org/10.1038/s41598-024-49923-3. Article 5521.
12. Das S, Panda SK, Sethy M, Giri PP, Nanda AK. (2024). *Comparative analysis of machine learning and deep learning models for classifying squamous epithelial cells of the cervix* (arXiv Preprint). https://arxiv.org/abs/2411.13535
13. Gruber K. New agreement aims to eliminate cervical cancer across the Americas. Lancet Oncol. 2025;26(5):355–6. https://doi.org/10.1016/S1470-2045(25)00126-3.
14. Jin Y, Ma J, Lian Y, Wang F. Cervical cytology screening using a fused deep learning architecture with attention mechanisms. Appl Soft Comput. 2024;166:112202. https://doi.org/10.1016/j.asoc.2024.112202.
15. Kadri AF, Babatunde RS, Olajide AT, Lawal OL, Mohammed SB, Isiaka OS, Ekundayo O, Babatunde AN. Logistics regression for prediction of breast cancer based on risk factor. FUW Trends Sci Technol J. 2025a;10(2):197–212.
16. Kadri AF, Mohammed SB, Ekundayo O, Tajudeen KO, Illori O, Popoola M, Isiaka OS, Babatunde AN. A modular CNN-based framework for real-time handwritten digit classification. Nigerian J Technological Dev. 2025b;22(4):25–40. https://doi.org/10.63746/njtd/v22i4.3840.
17. Kamzayeva N. Enhancing cervical cancer screening: New diagnostic strategies. Life. 2025;15(3):367. https://doi.org/10.3390/life15030367.
18. Karadag R, Yildirim T, Ucar A. Cervical cancer cell classification with deep learning on SIPaKMeD dataset. Biomed Signal Process Control. 2024;88:105653. https://doi.org/10.1016/j.bspc.2023.105653.
19. Khozaimi R, et al. Dense-UNet with metaheuristic optimisation for cervical cell segmentation. Comput Methods Programs Biomed. 2025;241:107832. https://doi.org/10.1016/j.cmpb.2024.107832.
20. Li Z, Liu Y, Xie F. A segmentation model to detect cervical lesions based on colposcopic images. Heliyon. 2023;9(9):e18451. https://doi.org/10.1016/j.heliyon.2023.e18451.
21. Lin H, Hu Y, Chen S, Yao J, Zhang L. (2018). *Fine-grained classification of cervical cells using morphological and appearance-based convolutional neural networks* (arXiv Preprint). https://arxiv.org/abs/1810.06058
22. Liu H, Zhang X, Chen L, Wang Y. Hybrid Swin Transformer and CNN ensembles for cervical cytology classification. Expert Syst Appl. 2024;238:121893. https://doi.org/10.1016/j.eswa.2023.121893.
23. Mosquera-Zamudio A, Cancino S, Cárdenas-Montoya G, García-Arteaga JD, Zambrano-Betancourt C, Parra-Medina R. Deep-learning approaches for cervical cytology nuclei segmentation in whole slide images. J Imaging. 2025;11(5):137. https://doi.org/10.3390/jimaging11050137.

24. Mulmule PV. Classification of cervical cytology overlapping cell images with transfer learning architectures. Biomedical Pharmacol J. 2022;15(1):147–55. https://doi.org/10.13005/bpj/2392.
25. Qian X, Li Y, Chen J. Stacking ensemble deep learning for cervical cytology classification. Comput Biol Med. 2023;158:106786. https://doi.org/10.1016/j.compbiomed.2023.106786.
26. Rahaman MM, Li C, Yao Y, Kulwa F, Wu X, Li X, Wang Q. DeepCervix: A deep learning-based framework for classification of cervical cells using hybrid deep feature fusion techniques. arXiv Preprint; 2021. https://arxiv.org/abs/2102.12191.
27. Rasheed A, Shirazi SH, Umar AI, Shahzad M, Yousaf W, Khan Z. Cervical cell nucleus segmentation through an improved U-Net architecture. PLoS ONE. 2023;18(10):e0283568. https://doi.org/10.1371/journal.pone.0283568.
28. Sabeena K, Ravishankar S, Pandurangan D. A hybrid model for efficient cervical cell classification. Comput Biol Med. 2022;146:105677. https://doi.org/10.1016/j.compbiomed.2022.105677.
29. Song Y, Zou J, Choi KS, Lei B, Qin J. Cell classification with worst-case boosting for intelligent cervical cancer screening. Med Image Anal. 2024;91:103014. https://doi.org/10.1016/j.media.2023.103014.
30. Tomaziu-Todosia ET, Anton G-I, Scripcariu I-S, Dumitrașcu I, Scripcariu DV, Balmus I-M, Ionescu C, Visternicu M, Socolov DG. Oxidative stress, inflammation, and antioxidant strategies in cervical cancer: A narrative review. Int J Mol Sci. 2025;26(10):4961. https://doi.org/10.3390/ijms26104961.
31. Wang H, Chen X, Zhang L, Liu J. Fully automatic segmentation method for cervical vertebral body detection. Digit Health. 2025;11:20552076251347695. https://doi.org/10.1177/20552076251347695.
32. Wubineh BZ, Rusiecki A, Halawa K. Classification of cervical cells from the Pap smear image using the RES_DCGAN data augmentation and ResNet50V2 with self-attention architecture. Neural Comput Appl. 2024;36:21801–15. https://doi.org/10.1007/s00521-024-10404-x.
33. Yang Y, Wang M, Ma L, Zhang X, Zhang K, Zhao X, Teng Q, Liu H. Cervical intervertebral disc segmentation based on multi-scale information fusion and its application. Electronics. 2024;13(2):432. https://doi.org/10.3390/electronics13020432.
34. Zhang Y, Ning C, Yang W. An automatic cervical cell classification model based on improved DenseNet121. Sci Rep. 2025;15(1):3240. https://doi.org/10.1038/s41598-025-87953-1.
35. Zhang Y, Zhu M, Luo Z. SegX: Improving interpretability of clinical image diagnosis with segmentation-based enhancement. arXiv Preprint; 2025. https://arxiv.org/abs/2502.10296.
36. Zhong YF, Liu X, Chen Z. A quantitative evaluation of a deep learning model of cervical spine MR image segmentation. Med Phys. 2024;51(2):14282. https://doi.org/10.1002/acm2.14282.

Publisher's note

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