Automated Segmentation of Cervical Cell Pap Smear Images for Malignancy Classification

Venkatesh Annasaheb Patil, R Nikhil Yadav, Ruthu H.M, Kiruthikaa P., Aneesh Rao, Meena, Dr. Ajey S.N.R
Department of Electronics and Communication Engineering
PES University, Bengaluru, India
Email: {venkatesh, nikhil, ruthu, kiruthikaa, aneesh, meena, ajey}@pes.edu

***Abstract*—Cervical cancer remains a major health concern globally, particularly in regions with limited access to healthcare infrastructure. Pap smear screening, although effective, still relies heavily on manual examination, which is time-consuming and susceptible to human error. In this paper, we propose an automated image segmentation pipeline for cervical cell Pap smear images that utilizes traditional image processing techniques, namely thresholding and morphological operations. The system segments critical structures such as the nucleus and cytoplasm, extracts key features, and evaluates segmentation quality using PSNR and SSIM metrics. The pipeline is designed for efficiency, reproducibility, and interpretability, making it suitable for real-time screening support and as a preprocessing tool for machine learning classifiers.**

***Index Terms*—Cervical cancer, Pap smear, Image segmentation, PSNR, SSIM, Morphological operations, OpenCV, Cytology, Medical imaging**

I. Introduction

Cervical cancer is one of the leading causes of cancer-related mortality among women globally, especially in low- and middle-income countries. Regular screening through Pap smear tests has significantly reduced mortality rates in regions with established healthcare programs. However, manual analysis of Pap smear images is labor-intensive and requires trained cytologists. The scarcity of expert personnel, especially in rural areas, highlights the urgent need for automated diagnostic tools.

The goal of automated analysis is to segment key structures such as the nucleus and cytoplasm in cervical cells, extract relevant morphological features, and assist in malignancy classification. A reliable segmentation method not only streamlines the screening workflow but also improves diagnostic consistency and reduces false positives or negatives.

This paper introduces a modular and interpretable segmentation pipeline developed using OpenCV in Python. The system uses grayscale conversion, intensity-based thresholding, and morphological operations to isolate regions of interest. Key features such as circularity, red-green intensity ratios, and centroid distances are calculated, and segmentation quality is assessed using PSNR and SSIM. Our method is lightweight, efficient, and adaptable for various imaging conditions.

II. Related Work

Over the past decade, various approaches have been proposed for cervical cell segmentation. Traditional methods include edge detection, histogram equalization, and morphological transformations. Alias et al. [1] conducted a comparative study evaluating segmentation methods using PSNR and SSIM, concluding that hybrid approaches yield better results in noisy or low-contrast datasets.

Zhang et al. [2] introduced a graph cut-based approach to cytoplasm segmentation, demonstrating high accuracy but computational overhead. Deep learning-based methods have also gained popularity, such as convolutional neural networks (CNNs) trained on labeled Pap smear datasets. While powerful, these models often lack interpretability and require large amounts of annotated data for training.

This paper focuses on traditional, rule-based methods that provide a balance between computational simplicity and practical accuracy. These methods are especially beneficial in resource-limited settings and are easier to validate and deploy in clinical workflows.

III. Proposed Methodology

The pipeline includes the following key stages:

A. Image Acquisition and Preprocessing

Pap smear images were collected in standard formats (JPEG, PNG or TIFF). These images included the classifications NILM (Negative for Intraepithelial Lesion or Malignancy), LSIL (Low-grade Squamous Intraepithelial Lesion), HSIL (High-grade Squamous Intraepithelial Lesion), and Carcinoma. These were then converted to grayscale. Further, histogram equalization improved contrast, while Gaussian blurring helped suppress high-frequency noise. These steps reduce computational complexity while preserving key structural details.

B. Thresholding and Morphological Processing

Two separate intensity thresholds were used:

- Dark threshold: to segment nuclei (intensity between 50-100).
- Medium threshold: to segment cytoplasm (intensity between 100-150).

Morphological operations, specifically closing (dilation followed by erosion), were applied to smoothen boundaries and eliminate internal gaps. These operations ensured cleaner segmentation masks.

C. Contour Detection

Contours of segmented regions were detected using OpenCV's contour detection algorithms, to identify closed regions that may represent individual cells or cellular structures. These contours were then filtered based on area to eliminate non-relevant segments such as noise or staining artifacts.

Color masks were generated by analyzing red and green intensity levels, which can represent different types of stains applied during cytology procedures. This step allows for mask overlays to isolate specific components, like cytoplasm (typically green-tinted) or nuclei (typically darker or reddish).

D. Feature Extraction and Ratio calculation

Following features were extracted from each segmented region:

- Area for identifying enlarged or shrunken cells.
- Perimeter for determining the complexity of cell shapes.
- Circularity C = Ratio of area to perimeter.
- Red-green ratio.
- Spatial Distance between centroids of nucleus and cytoplasm.

E. Statistical Summarization and Visualization

The dataset comprising extracted features were subjected to statistical analysis to detect patterns, abnormalities, and correlations. Descriptive statistics such as mean, median, interquartile range, standard deviation and advanced statistical methods such as normality tests to assess whether the red-green ratio follows a Gaussian distribution, Pearson correlation coefficients to examine the linear relationship between different features such as stained region areas etc were computed for each feature. The system automatically generates:

- Histograms to assess frequency distributions.
- Plots to visualize correlations.
- Plots to highlight statistical spread and outliers.

F. Image Quality Assessment

To verify that segmentation does not degrade diagnostic information, PSNR and SSIM were computed:

- **PSNR:** Ratio between the maximum possible power of a signal (the original image) and the power of noise (error introduced during processing).

$$\text{PSNR} = 10 \cdot \log_{10} \left(\frac{MAX_I^2}{\text{MSE}} \right) \tag{1}$$

  Where:
  - MAX_I is the maximum possible pixel value (255 for 8-bit grayscale images),
  - MSE is the Mean Squared Error between the original and segmented image.

  A higher PSNR value indicates lesser distortion and better image fidelity.

- **SSIM:** Evaluates perceptual similarity by accounting for structural, luminance, and contrast differences between the original and processed images.

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{2}$$

  Where:
  - μ_x, μ_y are mean intensities,
  - σ_x^2, σ_y^2 are variances,
  - σ_{xy} is the covariance between the images,
  - C_1 and C_2 are constants to stabilize division with weak denominators.

  SSIM values range from -1 to 1, with values closer to 1 representing high similarity.

IV. Experimental Results

The system was tested on categorized Pap smear images representing:

- NILM (Negative for Intraepithelial Lesion or Malignancy)
- LSIL (Low-grade Squamous Intraepithelial Lesion)
- HSIL (High-grade Squamous Intraepithelial Lesion)
- Carcinoma (Invasive cancer)

Each image underwent preprocessing, segmentation, and annotation. Visual results were saved with contours labeled, and feature values tabulated.

A. Feature Distribution Analysis

Box plots and histograms revealed distinctive patterns in circularity and red-green ratios across categories. For example, NILM cells showed high circularity and low red-to-green ratios, while carcinoma cells exhibited irregular shapes and elevated nucleus staining.

B. Sample Outputs

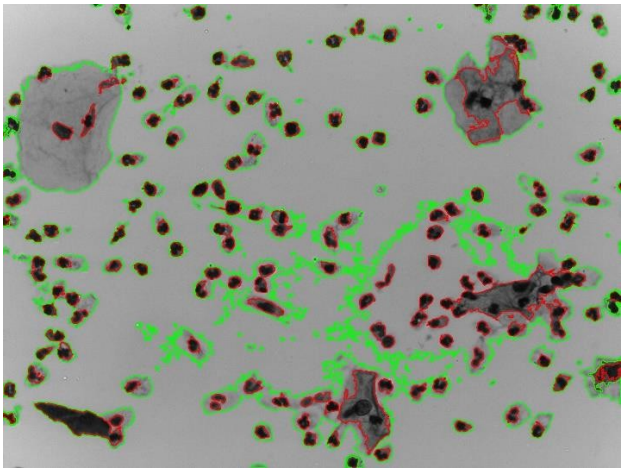

Fig. 1. Segmented NILM cell image with annotated nucleus and cytoplasm

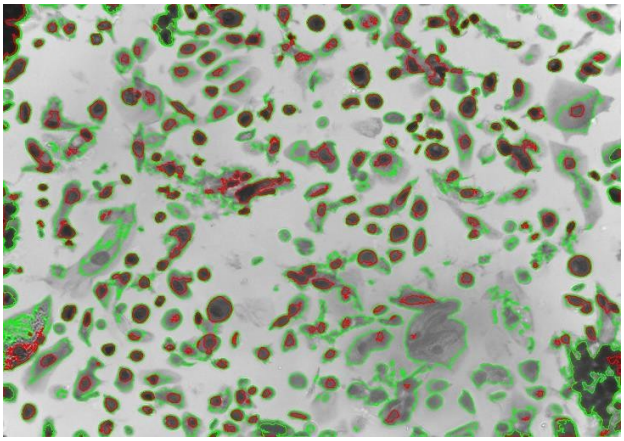

Fig. 2. Segmented HSIL cell image with annotated nucleus and cytoplasm

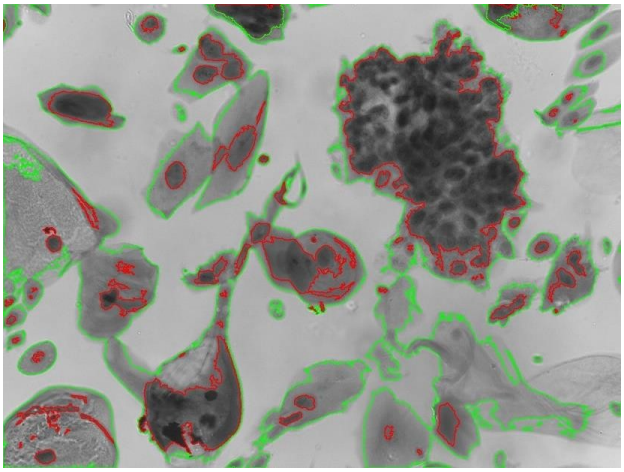

Fig. 3. Segmented Carcinoma image showing larger nuclei and distorted shape

V. Discussion

The segmentation pipeline provides a robust, interpretable method for preprocessing cytological images. It maintains high visual fidelity while accurately isolating biologically relevant structures. Compared to deep learning methods, our pipeline requires no training data and can be easily adapted to different datasets by tuning thresholds.

A limitation is that fixed thresholds may not generalize well to all staining variations. Future extensions could include adaptive thresholding using clustering techniques (e.g., Otsu's method) or hybrid ML-assisted segmentation for enhanced flexibility.

VI. Conclusion

This study proposes a reproducible and interpretable pipeline for the segmentation of cervical cell Pap smear images. Using intensity-based segmentation and morphological filtering, the system isolates regions of interest with high fidelity, evaluates image integrity using PSNR and SSIM, and extracts diagnostic features. The system can serve as a reliable preprocessing step for machine learning classifiers or as a stand-alone support tool in resource-limited clinical settings.

Acknowledgment

The authors express their sincere gratitude to Dr. Ajey S.N.R for his guidance, resources, and mentorship during the development of this project.

References

[1] Alias NA, et al. "Improvement method for cervical cancer detection," Oncol Res., 2022.

[2] Zhang L, et al. "Cytoplasm segmentation on cervical cell images," Biomed Mater Eng., 2014.

[3] Wang Z, et al. "Image quality assessment: from error visibility to structural similarity," IEEE TIP, 2004.

[4] Mudeng V, et al. "Prospects of Structural Similarity Index," Appl Sci., 2022.

[5] Widodo C, et al. "Medical image processing using Python and OpenCV," J. Phys.: Conf. Ser., 2020.