Report On

"Segmentation of Cervical Cell Pap Smear Images for Malignancy Classification"

Submitted in partial fulfilment of the requirements for the award of the degree of

Bachelor of Technology
in
Electronics & Communication Engineering

UE21EC390B – Capstone Project

Submitted by:

VENKATESH ANNASAHEB PATIL	**PES2UG21EC157**
R NIKHIL YADAV	**PES2UG21EC110**
RUTHU HM	**PES2UG21EC123**
KIRUTHIKAA.P	**PES2UG21EC100**
ANEESH RAO	**PES2UG21EC014**
MEENA	**PES2UG21EC081**

Under the guidance of
Dr. AJEY S.N.R.
Head Of Department - ECE
PES University

January – December 2024

DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING
FACULTY OF ENGINEERING
PES UNIVERSITY
(Established under Karnataka Act No. 16 of 2013)
Electronic City, Hosur Road, Bengaluru – 560 100, Karnataka, India

PES UNIVERSITY

(Established under Karnataka Act No. 16 of 2013)
Electronic City, Hosur Road, Bengaluru – 560 100, Karnataka, India

FACULTY OF ENGINEERING

CERTIFICATE

This is to certify that the dissertation entitled

‘Segmentation of Cervical Cell Pap Smear Images for Malignancy Classification’

is a Bonafide work carried out by

VENKATESH ANNASAHEB PATIL	**PES2UG21EC157**
R NIKHIL YADAV	**PES2UG21EC110**
RUTHU H.M.	**PES2UG21EC123**
KIRUTHIKAA .P	**PES2UG21EC100**
ANEESH RAO	**PES2UG21EC014**
MEENA	**PES2UG21EC081**

In partial fulfilment for the completion of seventh semester Capstone Project (UE21EC390B) in the Program of Study **-** Bachelor of Technology in Electronics and Communication Engineering under rules and regulations of PES University, Bengaluru during the period January – December 2024. It is certified that all corrections / suggestions indicated for internal assessment have been incorporated in the report. The dissertation has been approved as it satisfies the 7th semester academic requirements in respect of project work.

Signature	Signature
Dr. Ajey SNR	Prof. Nagarjuna Sadineni
Chairperson	Pro Vice Chancellor

External Viva

Name of Examiners **Signature with Date**

1. ------------------------------------- ---

2. ------------------------------------- --

DECLARATION

We, *Venkatesh Annasaheb Patil, R Nikhil Yadav, Ruthu H.M., Kiruthikaa P., Aneesh Rao, and Meena,* hereby declare that the report entitled, **‘Segmentation of Cervical Cell Pap Smear Images for Malignancy Classification’,** is an original work done by us under the guidance of **Dr. Ajey S.N.R**, Head of Department, ECE Department and is being submitted in partial fulfilment of the requirements for completion of 7th Semester course work in the Program of Study, B. Tech in Electronics and Communication Engineering.–

PLACE: Bengaluru

DATE:

NAME AND SIGNATURE OF THE CANDIDATES

1. VENKATESH ANNASAHEB PATIL

2. R NIKHIL YADAV

3. RUTHU H.M

4. KIRUTHIKAA. P

5. ANEESH RAO

6. MEENA

ABSTRACT

This project presents an automated pipeline for segmenting Pap smear images to aid in classifying cervical cell malignancy stages such as NILM, LSIL, HSIL, and Carcinoma. Using OpenCV-based image processing, the method applies intensity thresholding and morphological operations to accurately segment dark and medium-intensity regions. Centroid analysis offers spatial insights, while PSNR and SSIM metrics ensure quality. Annotated outputs enhance interpretability, and the system supports batch processing for large datasets. This segmentation framework serves as a vital preprocessing step for machine learning models and clinical tools in cervical cancer screening and early diagnosis.

ACKNOWLEDGEMENT

We express our sincere gratitude to **Prof. Jawahar D**, Chancellor - PES University**, Dr. J. Surya Prasad**, Vice Chancellor - PES University, and Prof. **Nagarjuna Sadineni**, Pro Vice Chancellor for providing us with various opportunities and for their constant encouragement and support.

We extend our heartfelt thanks to our project guide, **Dr. Ajey S. N. R**, Department of Electronics and Communication Engineering, PES University, for his invaluable guidance, assistance, and encouragement throughout the development of this UE21EC390B - Capstone Project. His insights and expertise played a crucial role in shaping this work. We are also grateful for his efforts in organizing, managing, and supporting us throughout the project process.

Once again, our sincere appreciation goes to **Dr. Ajey S. N. R**, Chairperson, Department of Electronics and Communication Engineering, PES University Electronic City Campus, for the continuous knowledge, support, and resources provided by the department.

Finally, we would like to extend our deepest gratitude to our families and friends for their unwavering encouragement and support. Their motivation has been a driving force in the successful completion of this project.

CONTENTS

List Of Figures

CHAPTER 1

PREAMBLE

1.1 Introduction

The field of medical imaging has seen rapid advancements in recent years, driven by the need for faster, more accurate, and scalable diagnostic tools. One of the most critical areas of focus is cervical cancer screening, where Pap smear tests play a pivotal role in the early detection of cellular abnormalities. These images, although rich in diagnostic value, require detailed manual analysis which is both time-consuming and prone to subjectivity. As a result, automating the segmentation and interpretation of cervical cell images has emerged as an important area of research.

Traditional image analysis techniques, such as manual thresholding and rule-based segmentation, have been used to process Pap smear images. While they offer simplicity and control, they often fall short in handling variations in staining, lighting, and cellular morphology. These limitations restrict their effectiveness in real-world screening applications, especially when dealing with large-scale datasets or inconsistent image quality.

1.2 Problem Statement

To segment and analyze cervical cell regions in Pap smear images using image processing techniques.

o Apply threshold-based segmentation and morphological operations to isolate regions of interest.

o Extract and compute spatial and structural features such as the nucleus-cytoplasm area ratios and centroid distances.

o Evaluate the quality and consistency of segmentation using image fidelity metrics like PSNR and SSIM.

CHAPTER 2

LITERATURE REVIEW

2.1 Improvement Method for Cervical Cancer Detection: A Comparative Analysis

Year of Publication: 2023

Authors: Nur Ain Alias, Wan Azani Mustafa, Mohd Aminudin Jamlos, Ahmed Alkhayyat, Khairul Shakir Ab Rahman, Rami Q. Malik

Introduction: This paper compares several image processing methods—such as histogram equalization, contrast enhancement, and edge detection—for improved cervical cell segmentation and cancer detection. The paper places particular emphasis on PSNR and SSIM for evaluating visual fidelity, along with clinical metrics like accuracy and sensitivity.

Experiment: Experiments were performed using the Herlev Pap smear dataset. Multiple segmentation techniques were implemented in MATLAB, and their outputs were evaluated using PSNR, SSIM, and performance metrics such as sensitivity and accuracy.

Conclusion: Among the methods tested, histogram equalization and edge-based segmentation achieved the highest SSIM and PSNR scores. The study suggests combining techniques for optimal performance, especially in low-contrast or noisy datasets.

2.2 Cytoplasm Segmentation on Cervical Cell Images Using Graph Cut-Based Approach

Year of Publication: 2013

Authors: Ling Zhang, Hui Kong, Chien Ting Chin, Tianfu Wang, and Siping Chen

Introduction: The study proposes a cytoplasm segmentation algorithm based on graph-cut optimization. By converting images to the LAB color space and applying multi-threshold Otsu segmentation, the algorithm effectively isolates cytoplasmic regions from overlapping cervical cells.

Experiment: Experiments were conducted on grayscale-converted Pap smear images. Segmentation was assessed using PSNR and SSIM alongside visual inspection.

Conclusion: The graph cut approach achieved over 93% accuracy and demonstrated strong PSNR and SSIM values. The method, while accurate, has limitations when dealing with densely overlapped cells, which could be mitigated by integrating region-growing techniques.

2.3 Image Quality Assessment: From Error Visibility to Structural Similarity

Year of Publication: 2004
Authors: Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli
Introduction: This seminal paper introduces the Structural Similarity Index Measure (SSIM) as an alternative to traditional image quality assessment tools like PSNR. SSIM considers perceptual aspects of image quality, evaluating luminance, contrast, and structural consistency, making it ideal for medical image evaluation.
Experiment: A series of image distortion tests were conducted, including compression and noise-based degradation. SSIM values were compared against human visual judgments to validate its superiority over PSNR.
Conclusion: SSIM outperformed PSNR in correlating with human perceptual quality, especially in scenarios involving structural distortion. The metric has since become widely adopted in various domains, including biomedical image processing.

2.4 Prospects of Structural Similarity Index for Medical Image Analysis

Year of Publication: 2022
Authors: Vicky Mudeng, Minseok Kim, Se-woon Choe
Introduction: This paper explores the application of SSIM and its derivatives in medical image analysis. It critiques the limitations of traditional metrics like PSNR in capturing perceptual degradation, especially in structurally sensitive medical images such as MRIs and cytology slides.

Experiment: The study includes an extensive review and evaluation of SSIM across multiple image modalities, including simulated distortions and compression scenarios.

Conclusion: SSIM consistently outperforms PSNR in reflecting perceptual quality in medical images. The authors advocate for domain-adapted SSIM models and stress its importance in applications such as cytological image segmentation where structural integrity is critical.

2.5 Image Segmentation using Morphological Operations

Year of Publication: 2015

Authors: Diya Chudasama, Tanvi Patel, Shubham Joshi, Ghanshyam I. Prajapati

Introduction: This paper proposes an image segmentation technique that combines a fuzzy logic-enhanced Canny edge detector with basic morphological operations (dilation, flood fill, and erosion). The approach aims to improve segmentation performance by enhancing edge clarity and then refining the segmented regions using morphological processing.

Experiment: The authors implemented the method on sample grayscale images. The fuzzy-Canny hybrid was first applied to highlight edges, followed by morphological dilation to close gaps, flood filling to eliminate internal holes, and erosion to thin boundaries. Visual results of each stage are provided, comparing original and segmented outputs.

Conclusion: The proposed two-phase segmentation approach successfully extracts object boundaries with improved accuracy and clarity. Morphological operations proved effective for refining results and eliminating noise, offering a simpler alternative to complex segmentation techniques.

2.6 Medical Image Processing Using Python and OpenCV

Year of Publication: 2020

Authors: C.E. Widodo, Kusworo Adi, Rahmat Gernowo

Introduction: This paper introduces a Python and OpenCV-based framework for processing medical images. The authors aim to simplify complex image analysis tasks commonly used in diagnostic imaging. The study focuses on essential preprocessing techniques such as

histogram equalization, smoothing (blurring), erosion, and dilation — all of which are pivotal in preparing raw medical images for more advanced stages like segmentation, feature extraction, and classification. The framework is developed with a practical approach, providing flexible tools for various types of biomedical image formats.

Experiment: The authors tested the framework on a set of grayscale medical images (e.g., X-rays, MRI scans). The images were first enhanced using histogram equalization to improve contrast. Then, Gaussian, and median blurring techniques were applied to reduce noise while preserving edges. Following noise suppression, the study implemented morphological operations — **erosion** to remove small artifacts and **dilation** to enhance cellular or

anatomical boundaries. The processed images were then passed through thresholding and contour detection routines to isolate specific regions of interest (ROIs), such as lesions or cellular components. Visual outputs were compared at each stage of preprocessing, showing significant improvement in clarity and segmentation readiness. The authors provided side-by-side comparisons between raw and processed images to demonstrate the effectiveness of each method.

Conclusion:

The paper concludes that OpenCV, when combined with Python, provides an accessible and efficient solution for preprocessing medical images. The implemented techniques — particularly histogram equalization and morphological operations — significantly enhance image quality and accuracy in subsequent analysis. These improvements lay the foundation for further cell detection, segmentation, and quantification, especially in pathology and microscopy-based workflows.

2.7 An Image-Based Flow Cytometric Approach to the Assessment of the Nucleus-to-Cytoplasm Ratio

Year of Publication: 2021

Authors: Michael J. Moore, Joseph A. Sebastian, Michael C. Kolios

Introduction: This paper demonstrates the use of image-based flow cytometry (IFC) to assess the nucleus-to-cytoplasm (N:C) ratio in both malignant and non-malignant cell lines. The study highlights the potential of IFC as a high-throughput method for evaluating cell malignancy, which is crucial for cancer diagnostics.

Experiment: The authors utilized IFC to quantify the N:C ratio in various cell lines and compared the ratios in malignant versus non-malignant cells. They found that malignant cells exhibited distinct N:C ratio patterns, which could serve as indicators of malignancy.

The study also highlights how IFC provides a quantitative and efficient approach for high-throughput analysis.

Conclusion: The paper concludes that IFC can be a powerful tool for quantifying the N:C ratio in large-scale cancer studies. It provides an efficient and reliable method for cancer diagnostics, offering the potential for better detection of malignancy in clinical settings.

CHAPTER 3

BACKGROUND AND CASE STUDY

3.1 Overview of Cervical Cancer

Cervical cancer remains a leading cause of cancer-related mortality among women, particularly in developing nations. It arises from abnormal growth of cells in the cervix, often initiated by persistent infection with high-risk types of Human Papillomavirus (HPV). The disease progression typically follows a well-defined path from low-grade lesions (LSIL) to high-grade lesions (HSIL) and eventually to carcinoma if undetected and untreated. The World Health Organization (WHO) estimates that over 600,000 new cases and more than 300,000 deaths occur globally each year due to cervical cancer, a significant percentage of which could be prevented with timely screening and diagnosis.

3.2 Importance of Early Detection

Early detection significantly increases the success rate of treatment, reducing both morbidity and mortality. Routine screening methods, such as the Pap smear test, play a crucial role in identifying precancerous changes. In many health programs, women between the ages of 21 and 65 are advised to undergo regular cervical screening. However, the effectiveness of these programs often hinges on the availability of trained cytologists and the consistency of their analysis.

3.3 The Pap Smear Test

The Pap smear, or Papanicolaou test, is a cytological screening procedure where cells collected from the cervix are examined microscopically to detect abnormalities. These abnormalities include changes in nuclear size, cytoplasmic texture, and the overall cell morphology. Based on these features, cells can be classified into diagnostic categories:

- **NILM (Negative for Intraepithelial Lesion or Malignancy)**: Indicates healthy, non-cancerous cervical cells.

- **LSIL (Low-grade Squamous Intraepithelial Lesion)**: Represents mild abnormalities often caused by HPV infection. These lesions may resolve on their own but can progress if left unchecked.
- **HSIL (High-grade Squamous Intraepithelial Lesion)**: Reflects more serious precancerous changes that are likely to develop into cancer if not treated.
- **Carcinoma**: Indicates the presence of invasive cervical cancer cells.

The test, though reliable, is labour-intensive and subject to human variability, especially when handling large volumes of samples.

3.4 Digital Cytology and Whole-Slide Imaging

With advancements in digital pathology, traditional glass slides can now be digitized using high-resolution scanners to produce whole-slide images (WSIs). These digital images facilitate remote diagnosis and open opportunities for automated analysis. Digital cytology enables the application of computational tools to assist cytologists in identifying and quantifying abnormal cells more efficiently.

3.5 Challenges in Manual Screening

Manual Pap smear screening poses multiple challenges:

- Subjectivity in interpretation
- Fatigue-related errors
- High inter- and intra-observer variability
- Time constraints in high-throughput environments

These limitations have spurred interest in developing computer-aided diagnostic (CAD) systems that can assist in the segmentation and classification of cervical cells.

3.6 Fundamentals of Image Processing

Image processing encompasses a set of techniques to enhance, analyse, and manipulate images for better interpretation. In medical imaging, this involves operations like:

- **Filtering**: Used to reduce noise and enhance features. Common filters include Gaussian (smooths image), Median (removes salt-and-pepper noise), and Bilateral (preserves edges).
- **Thresholding**: Separates foreground (objects) from background based on pixel intensity. Otsu's method, for instance, calculates an optimal threshold that minimizes intra-class variance.
- **Edge Detection**: Identifies boundaries of objects using operators like Sobel, Prewitt, and Canny. These methods detect changes in intensity that signify edges.
- **Morphological Operations**: Derived from mathematical morphology, these include dilation (expands bright regions), erosion (shrinks bright regions), opening (erosion followed by dilation), and closing (dilation followed by erosion). They are particularly useful in cleaning up segmented images and preserving structure.

These methods help isolate regions of interest (e.g., nuclei and cytoplasm) and prepare the images for higher-level analysis.

3.7 Segmentation in Medical Imaging

Segmentation refers to the process of partitioning an image into meaningful parts. In the context of Pap smear images, this typically involves identifying and separating cellular structures such as:

- **Nucleus**: Typically darker, circular/oval, and centrally located. Its shape and size are key indicators of malignancy.
- **Cytoplasm**: The lighter, irregular region surrounding the nucleus. Changes in its texture, shape, and staining are also diagnostic indicators.

Accurate segmentation is vital, as it directly affects feature extraction and downstream classification accuracy. Poor segmentation may lead to incorrect diagnoses or missed abnormalities.

3.8 Techniques for Cervical Cell Segmentation

Several image processing-based segmentation techniques have been employed in the analysis of Pap smear images:

- **Thresholding**: Segments images based on pixel intensity values. Otsu's method is particularly popular for its automatic thresholding capability.
- **Region Growing**: Initiates from a seed point and includes neighbouring pixels based on similarity criteria.
- **Watershed Algorithm**: Treats the image like a topographic map and separates regions based on ridges. It is effective but prone to over-segmentation.
- **Graph Cuts**: Models segmentation as a graph partitioning problem and finds the optimal cut that separates foreground and background.
- **Active Contours (Snakes)**: Uses energy-minimizing curves to detect object boundaries. They are particularly useful for irregular and overlapping shapes.

Each method has its advantages and limitations depending on the staining quality, image resolution, and cell overlap.

3.9 Evaluation of Segmentation Quality

The quality of segmentation must be assessed using objective metrics. Commonly used metrics in biomedical imaging include:

- **PSNR (Peak Signal-to-Noise Ratio)**: Indicates the quality of reconstructed images compared to the original. It is calculated as: PSNR= $10.Log_{10}(\frac{MAX_I^2}{MSE})$ where MAX_I is the maximum possible pixel value and MSE is the mean squared error.

- **SSIM (Structural Similarity Index Measure)**: Evaluates image similarity by comparing luminance, contrast, and structure. It ranges from -1 to 1, where 1 indicates perfect similarity.

- **Dice Coefficient and Jaccard Index**: When ground truth is available, these metrics quantify the overlap between predicted and actual segmented regions. The Dice score is given by: Dice=$\frac{2|A\cap B|}{|A|+|B|}$.

These metrics provide quantitative validation to ensure that segmented images retain their diagnostic integrity.

3.10 Role of Morphological Features in Diagnosis

Quantitative features derived from segmented images can assist in classifying cervical cells. These features include:

- **Area**: Total number of pixels inside a region; helps in assessing cell size.
- **Perimeter**: Length of the cell boundary; often correlates with irregularity.
- **Circularity**: Calculated as $\frac{4\,\pi\,.\,Area}{Perimeter^2}$, this measures how close the shape is to a perfect circle.
- **Centroid Distances**: Used to measure spatial relationships between cell components, such as nucleus-to-nucleus or nucleus-to-cytoplasm distance.
- **Intensity Ratios**: Compare pixel intensity distributions (e.g., red-green ratios) to highlight staining differences.

These features are crucial inputs to machine learning or rule-based diagnostic models.

3.11 Need for Automation in Screening Workflows

Given the rising incidence of cervical cancer and the limitations of manual screening, automation in cytological analysis is not just beneficial—it is essential. Automated pipelines that can batch-process images, segment cellular components, and provide quantifiable metrics offer consistency, scalability, and efficiency. These systems support cytologists by acting as a second reader, reducing diagnostic load, and improving throughput.

CHAPTER 4

DESIGN AND IMPLEMENTATION

4.1 Block Diagram Representation

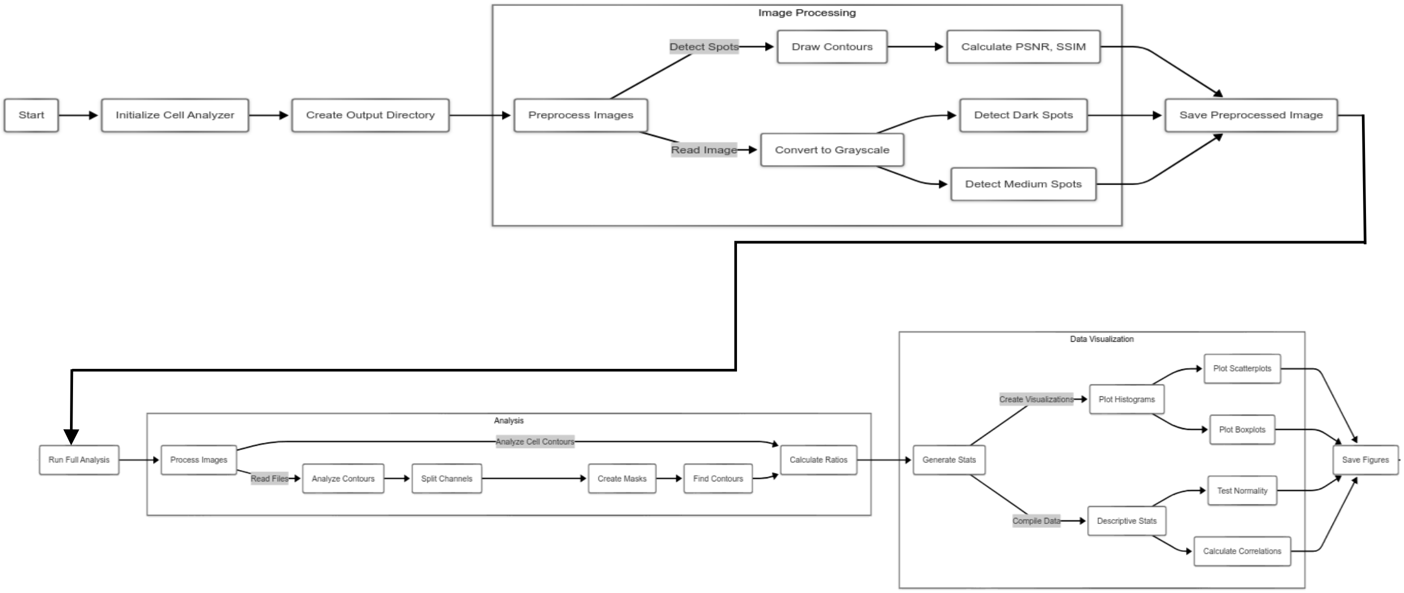


Fig 4.1.1: Block Diagram Representation of The Design Architecture

4.2 Implementation

4.2.1 System Setup and Dataset Acquisition

The first step in the pipeline involves importing a dataset of Pap smear images captured under a microscope. These images correspond to various stages of cervical cell development and may include classifications such as NILM (Negative for Intraepithelial Lesion or Malignancy), LSIL (Low-grade Squamous Intraepithelial Lesion), HSIL (High-grade Squamous Intraepithelial Lesion), and Carcinoma. To maintain structure and traceability, an output directory system is programmatically created. This ensures that all processed images, tabulated features, generated plots, and statistical summaries are stored with proper naming conventions and timestamps.

This step also includes scanning the dataset folder to identify all supported image formats, such as JPEG, PNG, or TIFF. This flexible approach allows users to handle diverse datasets sourced from different laboratories or research studies.

4.2.2 Preprocessing and Noise Filtering

Once the dataset is loaded, images undergo several preprocessing operations to enhance quality and accuracy for segmentation. Initially, the images are converted into grayscale to reduce computational complexity while preserving key structural details. Thresholding is then used to separate dark and medium-intensity pixel ranges, which helps differentiate between stained cell components such as nuclei and cytoplasm.

To further refine the binary masks generated through thresholding, morphological operations are applied—specifically closing operations to eliminate small gaps and noise. Additionally, individual color channels are analyzed separately to better isolate staining regions, particularly when color-based differentiation is essential. At this stage, image quality metrics such as Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) are also computed. These metrics help verify that preprocessing operations have preserved important visual details.

4.2.3 Saving and Preparing Segmented Images

Post preprocessing, each image with its respective contours and segmentation overlays is saved. These annotated outputs act as a visual validation step and allow for consistent comparisons across future experiments. Naming conventions include image identifiers, which assist in correlating processed results with original images.

Saving processed results not only reduces computation during future steps but also aids reproducibility. If any adjustments to threshold levels or morphological settings are required, previously saved outputs can be revisited without reprocessing the raw data.

4.2.4 Full Image Analysis and Mask Generation

The segmented and preprocessed images undergo deeper analysis to extract meaningful contours. Contour detection algorithms are applied to identify closed regions that may represent individual cells or cellular structures. These contours are filtered based on area to eliminate non-relevant segments such as noise or staining artifacts.

Color masks are 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).

4.2.5 Feature Extraction and Ratio Calculation

Following mask generation, features are extracted from each segmented region. These include:

- **Area**: Useful for identifying enlarged or shrunken cells.
- **Perimeter**: Helps determine the complexity of cell shapes.
- **Circularity**: Calculated using the ratio of area to perimeter, indicating shape regularity.
- **Red-Green Ratio**: Measures staining dominance and can be associated with cytological health.
- **Centroid Distance**: The spatial distance between stained components within a cell, offering insights into organization and clustering behavior.

Each feature is stored in structured tabular form, associating them with corresponding image identifiers and cell indices.

4.2.6 Statistical Summarization and Visualization

The dataset comprising extracted features is subjected to statistical analysis to detect patterns, abnormalities, and correlations. Descriptive statistics such as mean, median, interquartile range, and standard deviation are computed for each feature.

Advanced statistical methods are applied as well. For example, normality tests assess whether the red-green ratio follows a Gaussian distribution. Pearson correlation coefficients are computed to examine the linear relationship between different features such as stained region areas.

Visualization plays a crucial role in this phase. The system automatically generates:

- Histograms to assess frequency distributions.
- Scatter plots to visualize correlations.
- Box plots to highlight statistical spread and outliers.

4.3 Evaluation of Image Integrity

To ensure that the segmentation and annotation processes do not degrade or distort the original medical images, objective image quality metrics are employed after preprocessing. These metrics validate whether essential diagnostic information is preserved post-transformation.

Two primary metrics used are **Peak Signal-to-Noise Ratio (PSNR)** and **Structural Similarity Index Measure (SSIM)**.

Peak Signal-to-Noise Ratio (PSNR) quantifies the ratio between the maximum possible power of a signal (the original image) and the power of noise (error introduced during processing). It is calculated using:

$$PSNR = 10.Log_{10}(\frac{MAX_I^2}{MSE})$$

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.

Structural Similarity Index Measure (SSIM), on the other hand, evaluates perceptual similarity by accounting for structural, luminance, and contrast differences between the original and processed images. It is defined as:

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

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 C2C_2C2 are constants to stabilize division with weak denominators.

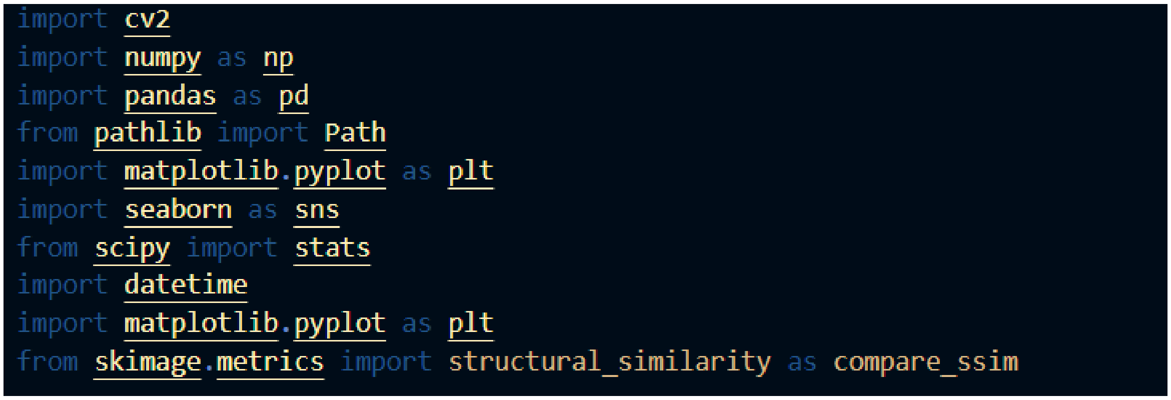
SSIM values range from -1 to 1, with values closer to 1 representing high similarity.

These metrics are calculated for every image processed, ensuring that the segmentation pipeline maintains both structural and perceptual quality. High PSNR and SSIM scores affirm that critical diagnostic features are not lost.

CHAPTER 5

CODING / ALGORITHM

5.1 Libraries Used



```
import cv2
import numpy as np
import pandas as pd
from pathlib import Path
import matplotlib.pyplot as plt
import seaborn as sns
from scipy import stats
import datetime
import matplotlib.pyplot as plt
from skimage.metrics import structural_similarity as compare_ssim
```

- **cv2 (OpenCV):** Used for reading, preprocessing, segmenting, and annotating Pap smear images through techniques like thresholding and contour detection.

- **numpy:** Supports numerical computations such as area, perimeter, and centroid calculations for segmented cellular regions.

- **pandas:** Stores and organizes extracted features (e.g., red-green ratios, circularity) into structured tables for statistical analysis and export.

- **pathlib:** Manages directory structures and file paths for saving processed images, plots, and result files systematically.

- **matplotlib.pyplot:** Generates visual plots including histograms, box plots, and scatter plots to visualize trends in cervical cell features.

- **seaborn:** Enhances data visualization with aesthetically appealing and informative plots to support statistical insights.

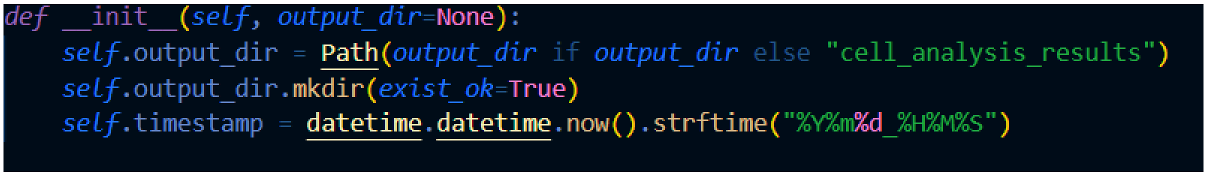
- **scipy.stats:** Performs statistical evaluations such as normality tests and correlation analysis on cell features extracted from images.

- **datetime:** Adds timestamps to output folders and files to maintain version control and track processing history.

- **skimage.metrics.structural_similarity (SSIM):** Quantifies visual similarity between original and processed images to ensure diagnostic integrity is preserved after segmentation.

5.2 Pseudo-Code Explanation

5.2.1 Image Folder Setup and Initialization



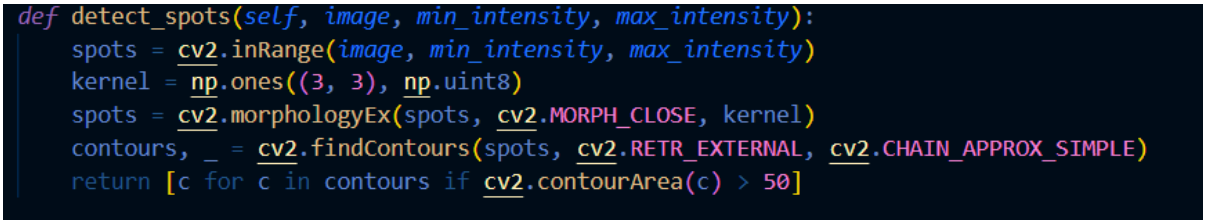
```
def __init__(self, output_dir=None):
    self.output_dir = Path(output_dir if output_dir else "cell_analysis_results")
    self.output_dir.mkdir(exist_ok=True)
    self.timestamp = datetime.datetime.now().strftime("%Y%m%d_%H%M%S")
```

• The system starts by specifying an input folder containing Pap smear images and an output folder for storing results.

• A timestamp is generated to uniquely tag result directories.

• Output folders are created for storing segmented images, feature tables, and visualizations.

5.2.2 Spot Detection Using Intensity Ranges



```
def detect_spots(self, image, min_intensity, max_intensity):
    spots = cv2.inRange(image, min_intensity, max_intensity)
    kernel = np.ones((3, 3), np.uint8)
    spots = cv2.morphologyEx(spots, cv2.MORPH_CLOSE, kernel)
    contours, _ = cv2.findContours(spots, cv2.RETR_EXTERNAL, cv2.CHAIN_APPROX_SIMPLE)
    return [c for c in contours if cv2.contourArea(c) > 50]
```

• Each grayscale image is read and examined to detect “dark” and “medium” stained regions.

• Intensity thresholds are applied to isolate pixel ranges:

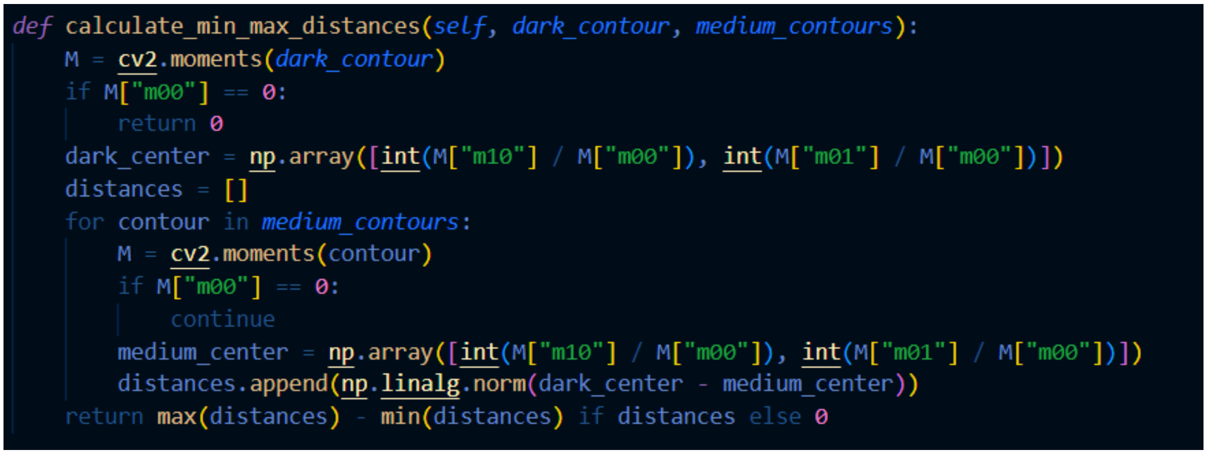
‣ Dark spots (e.g., nuclei) → pixel intensities between 50 and 100

‣ Medium spots (e.g., cytoplasm) → pixel intensities between 100 and 150

• Morphological closing is applied to clean the segmented regions.

• Contours are extracted and filtered based on area (>50 pixels) to remove noise.

5.2.3 Distance Analysis Between Cellular Regions



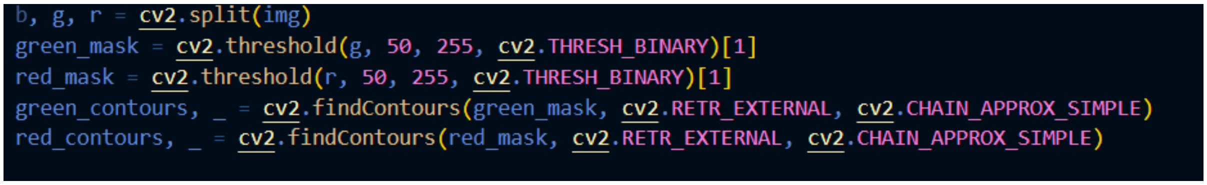
```python
def calculate_min_max_distances(self, dark_contour, medium_contours):
    M = cv2.moments(dark_contour)
    if M["m00"] == 0:
        return 0
    dark_center = np.array([int(M["m10"] / M["m00"]), int(M["m01"] / M["m00"])])
    distances = []
    for contour in medium_contours:
        M = cv2.moments(contour)
        if M["m00"] == 0:
            continue
        medium_center = np.array([int(M["m10"] / M["m00"]), int(M["m01"] / M["m00"])])
        distances.append(np.linalg.norm(dark_center - medium_center))
    return max(distances) - min(distances) if distances else 0
```

• For each dark spot, its centroid is calculated.

• The Euclidean distance from the dark spot to each medium-intensity region is computed.

• The difference between the farthest and nearest medium spot is recorded for spatial analysis.

5.2.4 Red-Green Channel-Based Segmentation



```python
b, g, r = cv2.split(img)
green_mask = cv2.threshold(g, 50, 255, cv2.THRESH_BINARY)[1]
red_mask = cv2.threshold(r, 50, 255, cv2.THRESH_BINARY)[1]
green_contours, _ = cv2.findContours(green_mask, cv2.RETR_EXTERNAL, cv2.CHAIN_APPROX_SIMPLE)
red_contours, _ = cv2.findContours(red_mask, cv2.RETR_EXTERNAL, cv2.CHAIN_APPROX_SIMPLE)
```

• The original image is split into red, green, and blue channels.

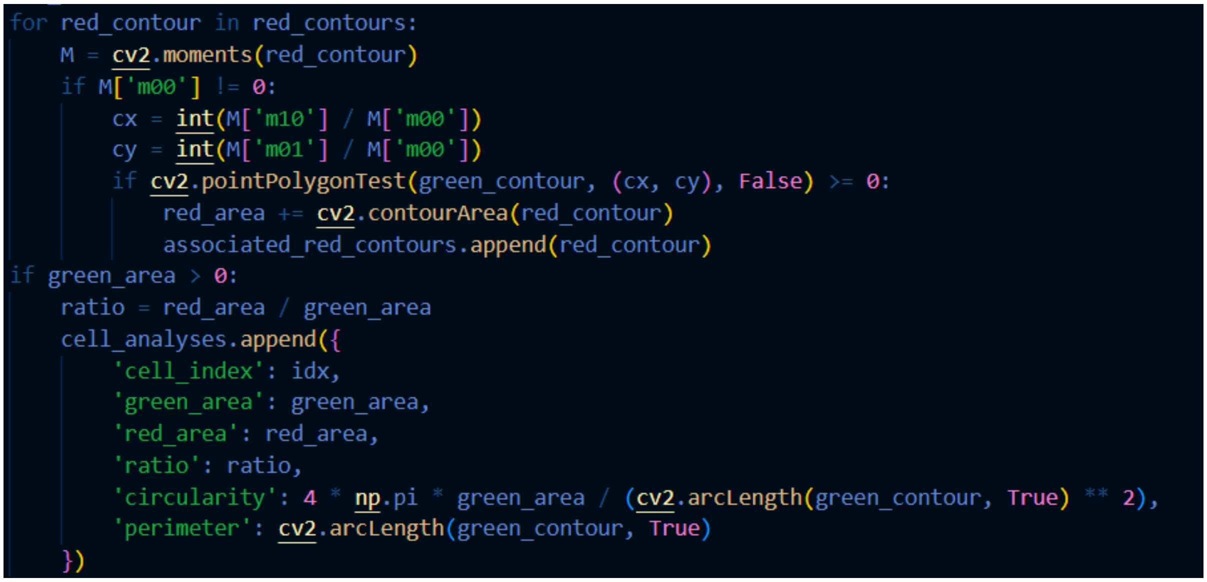
• Binary masks are created from the red and green channels using thresholding.

• Green contours are assumed to represent overall cell regions.

• Red contours within green contours are associated as possible nuclei or inner structures.

• For each green cell, red area, green area, perimeter, and shape metrics are computed.

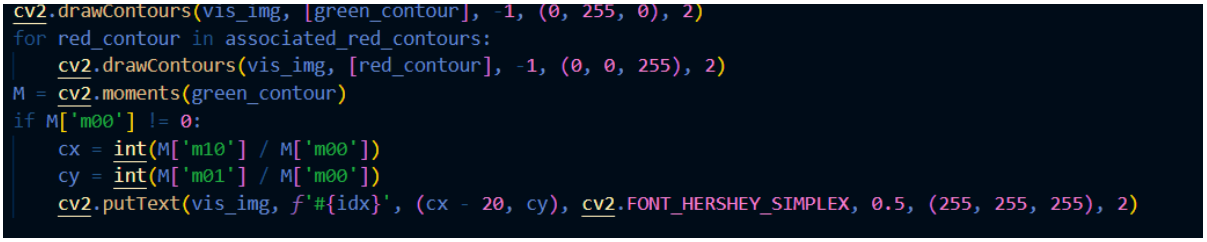
5.2.5 Feature Extraction and Ratio Calculation



```python
for red_contour in red_contours:
    M = cv2.moments(red_contour)
    if M['m00'] != 0:
        cx = int(M['m10'] / M['m00'])
        cy = int(M['m01'] / M['m00'])
        if cv2.pointPolygonTest(green_contour, (cx, cy), False) >= 0:
            red_area += cv2.contourArea(red_contour)
            associated_red_contours.append(red_contour)
if green_area > 0:
    ratio = red_area / green_area
    cell_analyses.append({
        'cell_index': idx,
        'green_area': green_area,
        'red_area': red_area,
        'ratio': ratio,
        'circularity': 4 * np.pi * green_area / (cv2.arcLength(green_contour, True) ** 2),
        'perimeter': cv2.arcLength(green_contour, True)
    })
```

- For each cell, the following features are extracted:
  - Green area
  - Red area
  - Red-to-green area ratio
  - Circularity
  - Perimeter length
- All features are stored in structured tables for each image.

5.2.6 Image Annotation and Visualization

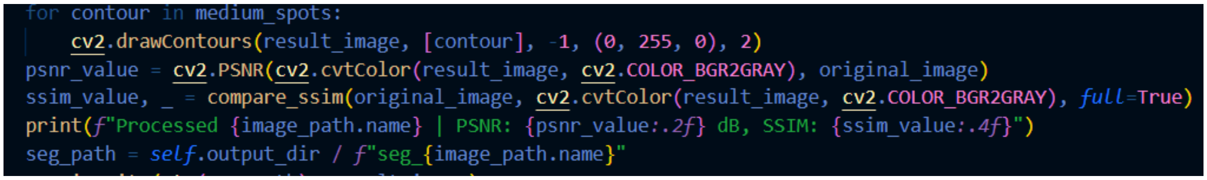


```python
cv2.drawContours(vis_img, [green_contour], -1, (0, 255, 0), 2)
for red_contour in associated_red_contours:
    cv2.drawContours(vis_img, [red_contour], -1, (0, 0, 255), 2)
M = cv2.moments(green_contour)
if M['m00'] != 0:
    cx = int(M['m10'] / M['m00'])
    cy = int(M['m01'] / M['m00'])
    cv2.putText(vis_img, f'#{idx}', (cx - 20, cy), cv2.FONT_HERSHEY_SIMPLEX, 0.5, (255, 255, 255), 2)
```

- All contours are drawn on the image with distinct colors:
  - Green for overall cells
  - Red for inner spots/nuclei

- Each green cell is labelled with an index number using centroid positioning.
- Annotated images are saved for visual validation.

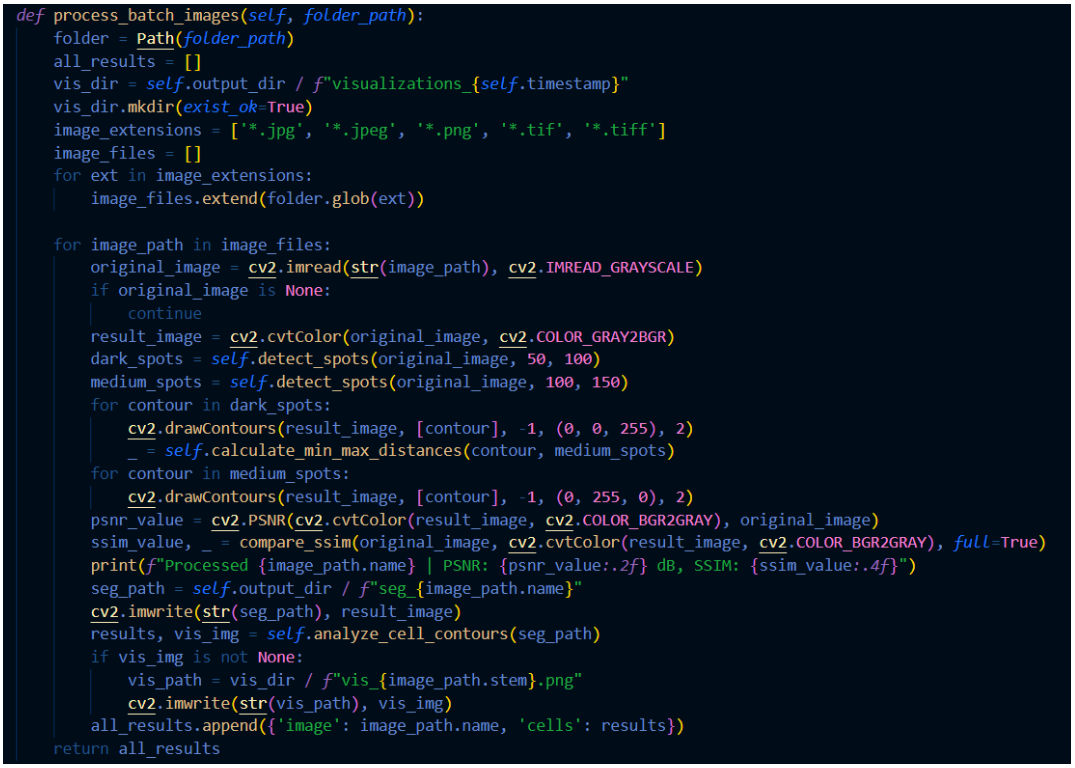
5.2.7 PSNR and SSIM Calculation for Image Quality Check



```
for contour in medium_spots:
    cv2.drawContours(result_image, [contour], -1, (0, 255, 0), 2)
psnr_value = cv2.PSNR(cv2.cvtColor(result_image, cv2.COLOR_BGR2GRAY), original_image)
ssim_value, _ = compare_ssim(original_image, cv2.cvtColor(result_image, cv2.COLOR_BGR2GRAY), full=True)
print(f"Processed {image_path.name} | PSNR: {psnr_value:.2f} dB, SSIM: {ssim_value:.4f}")
seg_path = self.output_dir / f"seg_{image_path.name}"
```

- PSNR and SSIM are computed between the original grayscale image and the processed color image.
- **PSNR** indicates the signal fidelity after annotation.
- **SSIM** evaluates the perceptual similarity across luminance, contrast, and structure.
- These metrics are logged for each image to ensure processing integrity.

5.2.8 Batch Image Processing and Data Aggregation

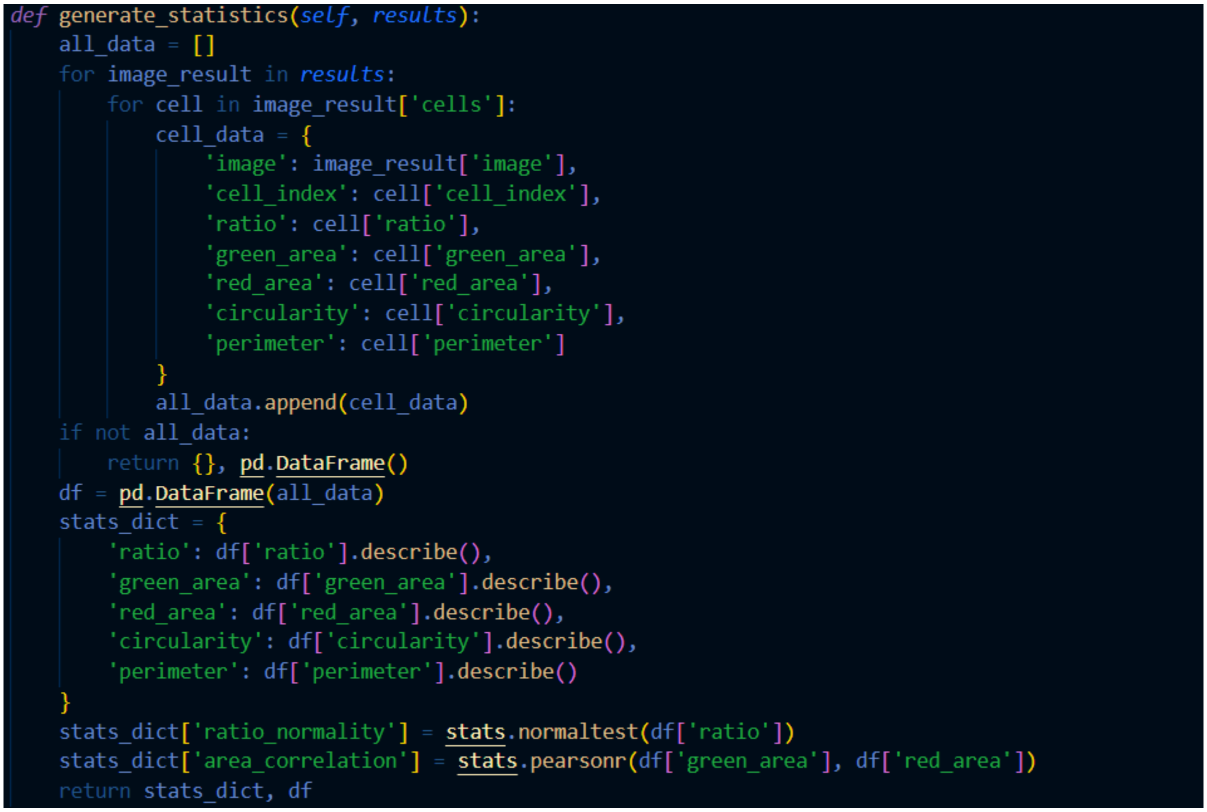


```
def process_batch_images(self, folder_path):
    folder = Path(folder_path)
    all_results = []
    vis_dir = self.output_dir / f"visualizations_{self.timestamp}"
    vis_dir.mkdir(exist_ok=True)
    image_extensions = ['*.jpg', '*.jpeg', '*.png', '*.tif', '*.tiff']
    image_files = []
    for ext in image_extensions:
        image_files.extend(folder.glob(ext))

    for image_path in image_files:
        original_image = cv2.imread(str(image_path), cv2.IMREAD_GRAYSCALE)
        if original_image is None:
            continue
        result_image = cv2.cvtColor(original_image, cv2.COLOR_GRAY2BGR)
        dark_spots = self.detect_spots(original_image, 50, 100)
        medium_spots = self.detect_spots(original_image, 100, 150)
        for contour in dark_spots:
            cv2.drawContours(result_image, [contour], -1, (0, 0, 255), 2)
            _ = self.calculate_min_max_distances(contour, medium_spots)
        for contour in medium_spots:
            cv2.drawContours(result_image, [contour], -1, (0, 255, 0), 2)
        psnr_value = cv2.PSNR(cv2.cvtColor(result_image, cv2.COLOR_BGR2GRAY), original_image)
        ssim_value, _ = compare_ssim(original_image, cv2.cvtColor(result_image, cv2.COLOR_BGR2GRAY), full=True)
        print(f"Processed {image_path.name} | PSNR: {psnr_value:.2f} dB, SSIM: {ssim_value:.4f}")
        seg_path = self.output_dir / f"seg_{image_path.name}"
        cv2.imwrite(str(seg_path), result_image)
        results, vis_img = self.analyze_cell_contours(seg_path)
        if vis_img is not None:
            vis_path = vis_dir / f"vis_{image_path.stem}.png"
            cv2.imwrite(str(vis_path), vis_img)
        all_results.append({'image': image_path.name, 'cells': results})
    return all_results
```

- All images in the input folder are processed one by one.
- Segmented results and annotations are saved for each file.
- Feature data from all images is aggregated into a single DataFrame.

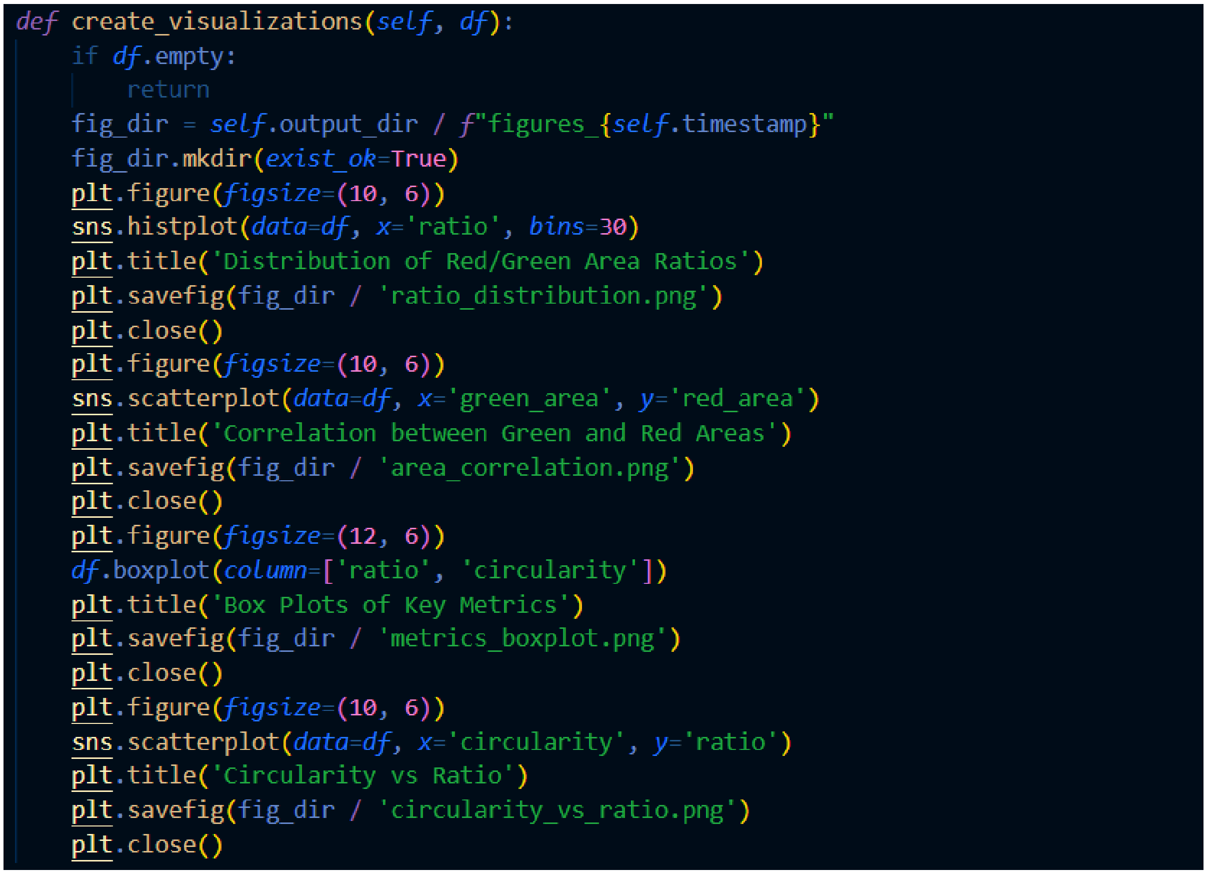
5.2.9 Statistical Analysis of Cell Features



```python
def generate_statistics(self, results):
    all_data = []
    for image_result in results:
        for cell in image_result['cells']:
            cell_data = {
                'image': image_result['image'],
                'cell_index': cell['cell_index'],
                'ratio': cell['ratio'],
                'green_area': cell['green_area'],
                'red_area': cell['red_area'],
                'circularity': cell['circularity'],
                'perimeter': cell['perimeter']
            }
            all_data.append(cell_data)
    if not all_data:
        return {}, pd.DataFrame()
    df = pd.DataFrame(all_data)
    stats_dict = {
        'ratio': df['ratio'].describe(),
        'green_area': df['green_area'].describe(),
        'red_area': df['red_area'].describe(),
        'circularity': df['circularity'].describe(),
        'perimeter': df['perimeter'].describe()
    }
    stats_dict['ratio_normality'] = stats.normaltest(df['ratio'])
    stats_dict['area_correlation'] = stats.pearsonr(df['green_area'], df['red_area'])
    return stats_dict, df
```

- Descriptive statistics are computed for all features.
- A normality test is performed on the red-green ratio to assess distribution shape.
- Pearson correlation is used to examine the relationship between red and green areas.

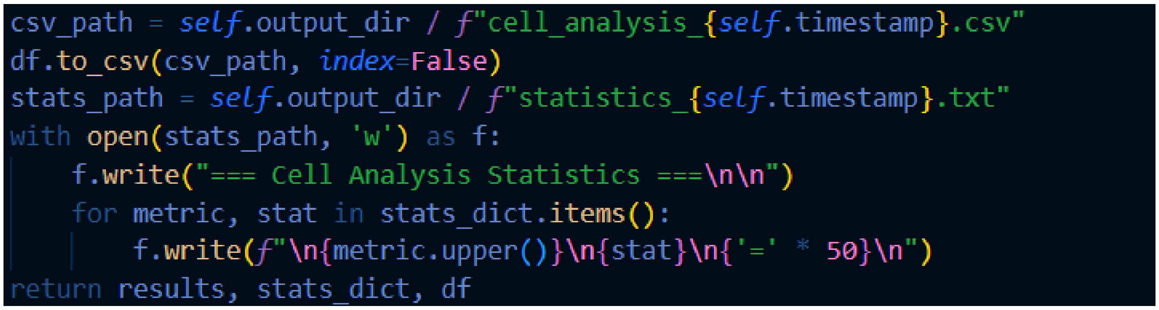
5.2.10 Visualization of Statistical Insights



```
def create_visualizations(self, df):
    if df.empty:
        return
    fig_dir = self.output_dir / f"figures_{self.timestamp}"
    fig_dir.mkdir(exist_ok=True)
    plt.figure(figsize=(10, 6))
    sns.histplot(data=df, x='ratio', bins=30)
    plt.title('Distribution of Red/Green Area Ratios')
    plt.savefig(fig_dir / 'ratio_distribution.png')
    plt.close()
    plt.figure(figsize=(10, 6))
    sns.scatterplot(data=df, x='green_area', y='red_area')
    plt.title('Correlation between Green and Red Areas')
    plt.savefig(fig_dir / 'area_correlation.png')
    plt.close()
    plt.figure(figsize=(12, 6))
    df.boxplot(column=['ratio', 'circularity'])
    plt.title('Box Plots of Key Metrics')
    plt.savefig(fig_dir / 'metrics_boxplot.png')
    plt.close()
    plt.figure(figsize=(10, 6))
    sns.scatterplot(data=df, x='circularity', y='ratio')
    plt.title('Circularity vs Ratio')
    plt.savefig(fig_dir / 'circularity_vs_ratio.png')
    plt.close()
```

- Histograms are plotted for red-green ratio distribution.
- Scatter plots show relationships (e.g., green vs. red area, circularity vs. ratio).
- Box plots are generated to examine variation and outliers.
- All plots are saved in the output directory with timestamped filenames.

5.2.11 Export of Results and Report Generation



```python
csv_path = self.output_dir / f"cell_analysis_{self.timestamp}.csv"
df.to_csv(csv_path, index=False)
stats_path = self.output_dir / f"statistics_{self.timestamp}.txt"
with open(stats_path, 'w') as f:
    f.write("=== Cell Analysis Statistics ===\n\n")
    for metric, stat in stats_dict.items():
        f.write(f"\n{metric.upper()}\n{stat}\n{'=' * 50}\n")
return results, stats_dict, df
```

- All feature tables are exported as `.csv` files.
- Statistical summaries are saved as `.txt` reports.
- Visual plots are stored in organized folders for easy reference.

CHAPTER 6

RESULTS AND VERIFICATION

6.1 NILM Outputs

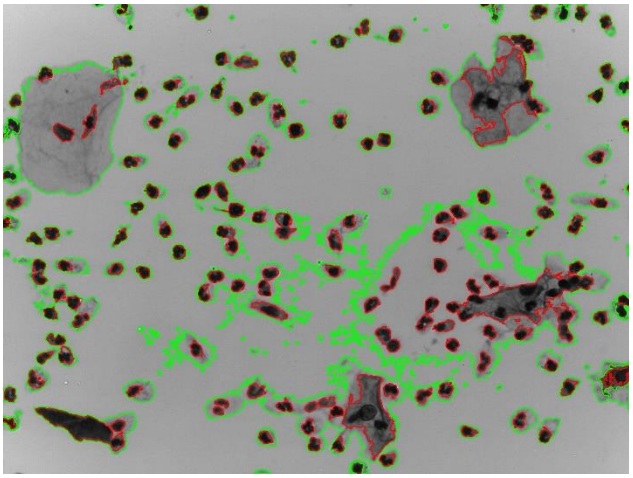


Fig. 6.1.1 Final segmented output for NILM pap smear cells.

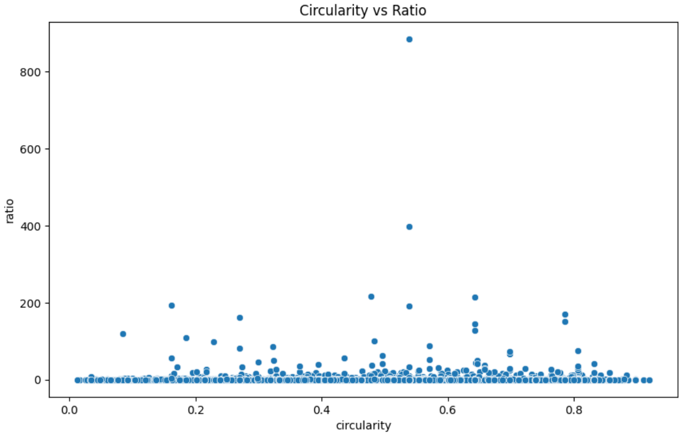


Fig 6.1.2 Circularity vs Ratio (of the nucleus and cytoplasm) plot for analysis of the individual cells as a statistic.

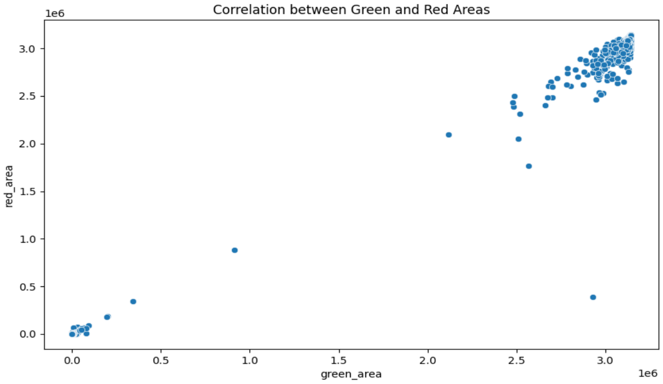


Fig 6.1.3 Correlation between the area of the nucleus and the area of the cytoplasm.

6.2 HSIL Outputs

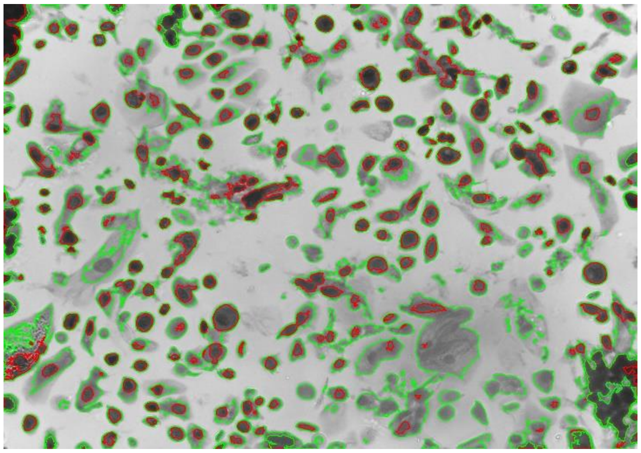


Fig 6.2.1 Final segmented output for HSIL pap smear cells.

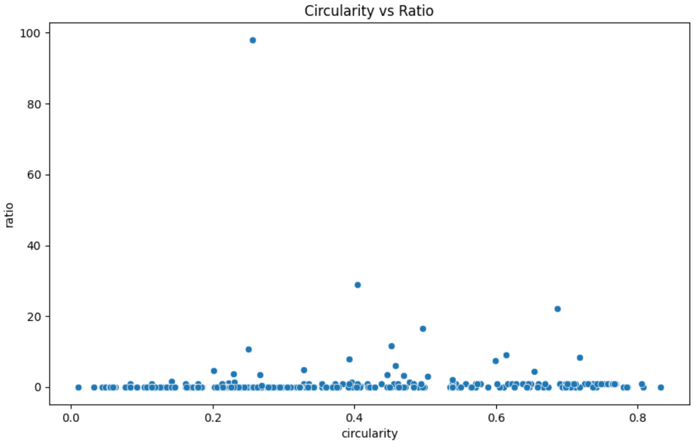


Fig 6.2.2 Circularity vs Ratio (of the nucleus and cytoplasm) plot for analysis of the individual cells as a statistic.

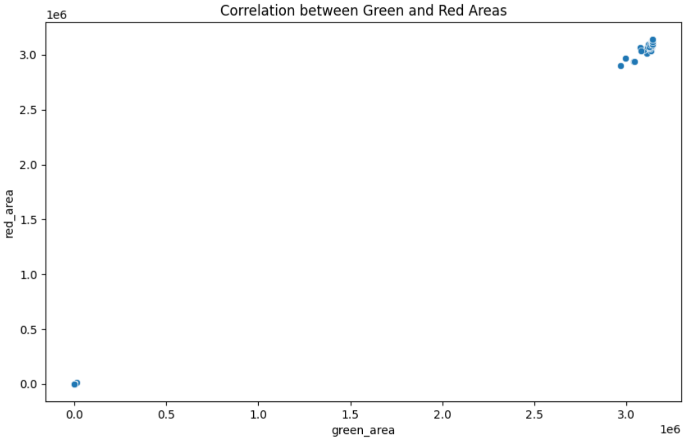


Fig 6.2.3 Correlation between the area of the nucleus and the area of the cytoplasm.

6.3 Carcinoma Outputs

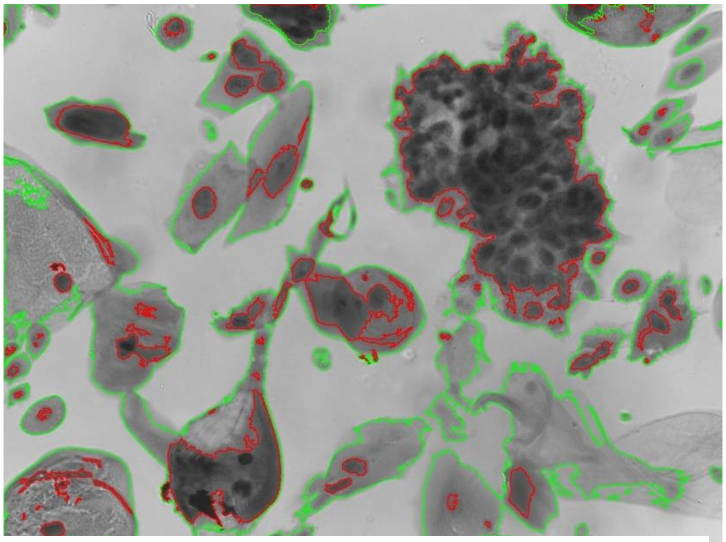


Fig 6.3.1 Final segmented output for Carcinoma pap smear cells. It is visible that there are irregular clustering and larger nucleus concentration.

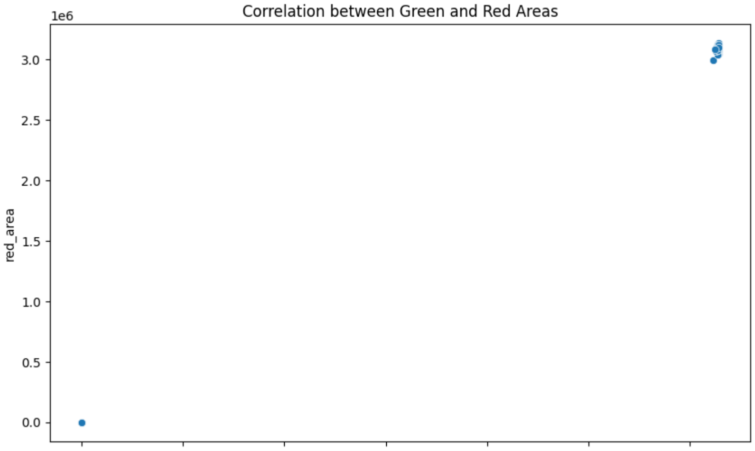


Fig 6.3.2 Correlation between the area of the nucleus and the area of the cytoplasm. Evident that there are undesirable irregularities when compared to NILM (non-cancerous).

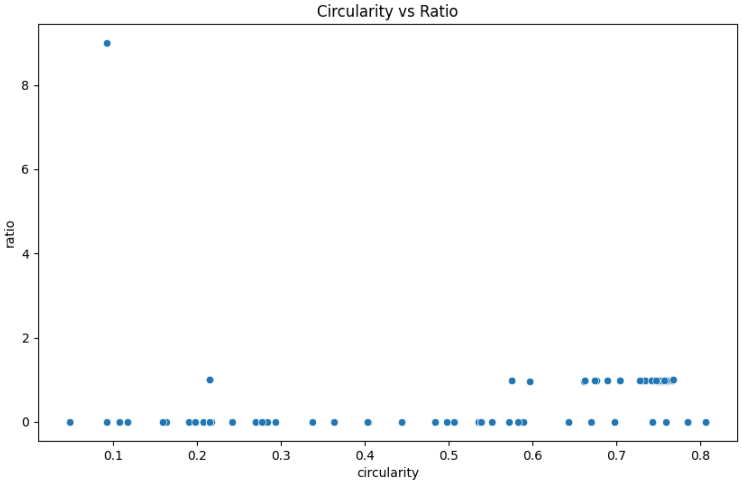


Fig 6.3.3 Circularity vs Ratio (of the nucleus and cytoplasm) plot for analysis of the individual cells as a statistic. This graph suggest the cells are more irregular and “non-circular”.

CHAPTER 7

CONCLUSION AND FUTURE WORK

7.1 Conclusions

This project presented a comprehensive, rule-based pipeline for the analysis of cervical cytology images, specifically Pap smear slides. By combining intensity-based segmentation, morphological operations, and contour detection techniques, the system successfully identified and annotated biologically significant regions such as nuclei and cytoplasm.

Quantitative metrics like **red-to-green area ratio**, **circularity**, and **perimeter** were extracted for each segmented cell, allowing for statistical characterization of cellular morphology. Image quality was rigorously evaluated using **PSNR** and **SSIM**, ensuring that the structural and perceptual integrity of processed images remained intact.

Furthermore, the pipeline supports **batch processing**, automated visualization, and output archiving, making it scalable and reproducible for real-world clinical or research applications. This foundational tool can act as a preprocessing module for machine learning-based classification models or aid in expert reviews.

In summary, the proposed methodology demonstrates:

- High interpretability and transparency in image analysis.
- Effective isolation of stained cellular regions.
- Robust statistical and visual insight into morphological variations.
- Preservation of diagnostic content post-processing.

7.2 Future Work

While the current pipeline effectively segments and analyzes Pap smear images using traditional image processing, several opportunities exist to extend its capabilities in future work:

Integration with AI/ML Models
The extracted features can be fed into supervised learning algorithms such as Support Vector Machines or Random Forests to automate the classification of malignancy stages. Deep learning models like CNNs can also be explored for end-to-end segmentation and diagnosis.

Advanced Feature Engineering

Future implementations could incorporate more sophisticated features such as nucleus-to-cytoplasm ratio (N/C ratio), texture descriptors, or spatial distribution metrics to improve diagnostic relevance.

Adaptive Segmentation Techniques

Adaptive thresholding or machine learning–guided segmentation could replace fixed intensity thresholds, improving robustness across varied staining conditions and imaging devices.

User Interface or Web Deployment

Developing a simple GUI or web-based tool would make the system accessible to clinicians and researchers without programming knowledge.

Validation with Expert Feedback

Collaborating with pathologists for manual validation would help refine accuracy and tailor the system for clinical deployment.

REFERENCES

[1] Alias NA, Mustafa WA, Jamlos MA, Alkhayyat A, Rahman KSA, Q Malik R. Improvement method for cervical cancer detection: A comparative analysis. Oncol Res. 2022 Oct 10;29(5):365-376. doi: 10.32604/or.2022.025897. PMID: 37305159; PMCID: PMC10208041.

[2] Zhang L, Kong H, Chin CT, Wang T, Chen S. Cytoplasm segmentation on cervical cell images using graph cut-based approach. Biomed Mater Eng. 2014;24(1):1125-31. doi: 10.3233/BME-130912. PMID: 24212005.

[3] Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process. 2004 Apr;13(4):600-12. doi: 10.1109/tip.2003.819861. PMID: 15376593.

[4] Mudeng V, Kim M, Choe S-w. Prospects of Structural Similarity Index for Medical Image Analysis. *Applied Sciences*. 2022; 12(8):3754. https://doi.org/10.3390/app12083754 .

[5] Chudasama, Diya & Patel, Tanvi & Joshi, Shubham & Prajapati, Ghanshyam. (2015). Image Segmentation using Morphological Operations. International Journal of Computer Applications. 117. 16-19. 10.5120/20654-3197.

[6] Widodo, C & Adi, Kusworo & Gernowo, Rahmat. (2020). Medical image processing using python and open cv. Journal of Physics: Conference Series. 1524. 012003. 10.1088/1742-6596/1524/1/012003.

[7] Sebastian JA, Moore MJ, Berndl ESL, Kolios MC. An image-based flow cytometric approach to the assessment of the nucleus-to-cytoplasm ratio. PLoS One. 2021 Jun 24;16(6):e0253439. doi: 10.1371/journal.pone.0253439. PMID: 34166419; PMCID: PMC8224973.