Acta Univ. Sapientiae Informatica **15,** 2 (2023) 306–329

DOI: 10.2478/ausi-2023-0019

AnnoCerv: A new dataset for feature-driven and image-based automated colposcopy analysis

Dorina Adelina MINCIUNĂ
University of Medicine and Pharmacy
Iași, Romania

Demetra Gabriela SOCOLOV
University of Medicine and Pharmacy
Iași, Romania

Attila SZŐCS ✉
Ascorb Research S.R.L.
Târgu Mureș, Romania
email: szocsatti@gmail.com

Doina IVANOV
University of Medicine and Pharmacy
Iași, Romania

Tudor GÎSCĂ
University of Medicine and Pharmacy
Iași, Romania

Valentin NECHIFOR
University of Medicine and Pharmacy
Iași, Romania

Sándor BUDAI
Cattus Distribution S.R.L.
Târgu Mureș, Romania

Attila GÁL
Cattus Distribution S.R.L.
Târgu Mureș, Romania

Ákos BÁLINT
Cattus Distribution S.R.L.
Târgu Mureș, Romania

Răzvan SOCOLOV
University of Medicine and Pharmacy
Iași, Romania

David ICLANZAN
Sapientia Hungarian University of Transylvania,
Târgu Mureș, Romania
ORCID: 0000-0003-2587-9106

Key words and phrases: colposcopy imaging, cervical cancer diagnosis, lesion segmentation, automated image analysis

Abstract. Colposcopy imaging is pivotal in cervical cancer diagnosis, a major health concern for women. The computational challenge lies in accurate lesion recognition. A significant hindrance for many existing machine learning solutions is the scarcity of comprehensive training datasets.

To reduce this gap, we present AnnoCerv: a comprehensive dataset tailored for feature-driven and image-based colposcopy analysis. Distinctively, AnnoCerv include detailed segmentations, expert-backed colposcopic annotations and Swede scores, and a wide image variety including acetic acid, iodine, and green-filtered captures. This rich dataset supports the training of models for classifying and segmenting low-grade squamous intraepithelial lesions, detecting high-grade lesions, aiding colposcopy-guided biopsies, and predicting Swede scores – a crucial metric for medical assessments and treatment strategies.

To further assist researchers, our release includes code that demonstrates data handling and processing and exemplifies a simple feature extraction and classification technique.

1 Introduction

Cervical cancer, characterized by a malignant tumor in the cervix, ranks as the fourth most prevalent cancer in women worldwide [19]. It accounts for approximately 6.6% of all female cancer cases due to its high incidence rate [19]. A critical concern is the absence of symptoms in the early stages, leading to a notably high mortality rate. According to the World Health Organization, there were an estimated 604000 new cases and 342000 deaths in 2020 [27]. Distressingly, around 90% of these instances were in low- and middle-income nations [27]. The key to combating this disease lies in the timely detection of precancerous lesions, early diagnosis, and prompt treatment. In this context, colposcopy emerges as a pivotal tool, significantly enhancing the cervical cancer detection rate and serving as an effective screening method for precancerous lesions [21, 24, 28].

Used primarily as a follow-up to abnormal Pap smear results, colposcopy provides a magnified view of the cervix, enabling healthcare providers to pinpoint potential areas of concern. This procedure aids in detecting and diagnosing various cervical issues, including cervical dysplasia, HPV infections, and inflammation [21, 24, 28]. By discerning the gravity and reach of these abnormalities, practitioners can make informed decisions. For instance, if anomalies are spotted during a colposcopy, a biopsy might be conducted. Furthermore, colposcopy is instrumental in monitoring treatment effectiveness for cervical abnormalities. After certain treatments, like tissue removal, consistent colposcopy exams ensure the healing process is on track and no new abnormalities arise. In scenarios demanding a more intensive treatment approach, colposcopy can guide surgical interventions, such as the loop electrosurgical excision procedure (LEEP) or cold knife cone biopsy [21, 24, 28].

Recognizing the pivotal role of colposcopy images in the diagnosis of cervical cancer, it is imperative to emphasize the significance of image quality for accurate analysis, particularly precancerous cervical lesions [11]. The need for high-quality imagery is amplified in telemedicine discussions among multiple doctors. Given the potential impact of variables - such as camera angles, lighting and shaking - on image quality, defects such as low contrast and distortion can compromise diagnosis precision [11].

Extensive research confirms high-risk human papillomavirus infection as a primary cause of cervical cancer [12, 17, 8, 25]. Early screening, when paired with HPV testing and cytology, has the potential to identify 80.7–98.7% of cervical intraepithelial neoplasia [26, 3]. Colposcopy-guided biopsies are the gold standard for detecting cervical cancer and its precancerous lesions. However, the precision of diagnosis can be influenced by various factors, from the expertise of the gynecologist to the woman's menstrual status. In particular, even for experienced gynecologists, the sensitivity of colposcopy for identifying cancerous lesions ranges from 81.4% to 95.7%, with a specificity between 34.2% and 69% [7, 23, 22]. Consequently, improving colposcopy precision is becoming a priority in the management of intraepithelial cervical neoplasia.

In contemporary medicine, artificial intelligence (AI) and deep learning have carved a niche, enabling efficient analysis of vast clinical data. Recent findings highlight the utility of medical AI and computer-assisted diagnosis in identifying cancerous lesions, leveraging deep learning and medical image processing techniques. Studies spanning optical tomography [18], radiology [14], computerized tomography [9], colonoscopy [2], and morphopathology [10] suggest that with ample training data, machine learning can rival or even surpass clinicians in diagnostic accuracy.

Historically, Acosta et al. [1] employed the K-NN algorithm to discern normal from abnormal cervical tissues, achieving 71% sensitivity and 59% specificity. Asiedu et al. [4] reported 81.3% sensitivity and 78.6% accuracy in distinguishing between cervical neoplasia and normal tissues. Liming Hu et al.'s [15] seven-year cohort study trained a deep learning algorithm on colposcopy images, achieving higher accuracy than the Pap smear. Additionally, Bing Bai et al. [5] in 2018 used the K-means algorithm for automatic cervical region segmentation. Deep learning methods, with their capacity to autonomously extract pertinent features from training data, underscore their value alongside conventional diagnostic techniques. However, a pertinent challenge remains: medical image datasets are often limited, constraining training capabilities.

Many studies keep their image datasets private. As a result, only a select few databases are available for developers [16, 30, 20, 13, 29]. Notably, the existing datasets primarily consist of acetic acid images. In light of this, there is an urgent need to develop or expand datasets, aiming to incorporate a diverse range of images, including acetic acid, iodine, and green-filter types.

In our study, we amassed a collection of colposcopy images. These images were meticulously segmented and annotated by specialists to distinctly visualize both healthy and pathological changes in cervical tissue. What sets this curated collection apart is its inclusion of acetic acid images, iodine images, and green-filtered images. This comprehensive dataset is now available for training machine learning models, aiding in the automatic classification and segmentation of low-grade squamous intraepithelial lesions (LSIL), detection of high-grade squamous intraepithelial lesions (HSIL), and assisting with colposcopy-guided biopsies. All curated data were cross-referenced with gynecological evaluations based on the patients' medical record.

2 Materials and methods

2.1 Comprehensive assessment in colposcopy: techniques and criteria

Colposcopy, a diagnostic procedure employed primarily in gynecological examinations, relies heavily on the discerning observation of cervical tissues to detect anomalies and potential malignancies. This procedure utilizes different techniques and criteria, each tailored to accentuate specific aspects of cervical tissue and enhance diagnostic precision. In the subsequent sub-sections, we will delve into the principal components and characteristics pivotal to colposcopy images, understand the significance of the Swede score evaluation as a diagnostic tool, and shed light on the transformation zone's classification methodology. Together, these criteria and methods offer a comprehensive overview, enabling a nuanced understanding of colposcopic examinations and the annotated images in the dataset.

2.1.1 Key elements and features of colposcopy images

The accurate diagnosis of cervical neoplasia using colposcopy is contingent on four primary features:

1. Intensity of Aceto-whitening: This refers to the color tone variations seen in the cervix upon application of acetic acid.
2. Demarcation and Surface Contour of Aceto-white Areas: This encompasses the clarity and texture of the white regions appearing after the acetic acid application.
3. Vascular Features: The visibility of blood vessels provides insights into the health of the cervical tissue.
4. Iodine-Induced Color Changes: Observing how the cervix responds to iodine application can give vital diagnostic clues.

Additional diagnostic considerations include:

- Anomalies in the transformation zone can be indicative of neoplasia.
- Expert gynecologists can differentiate between low-grade cervical intraepithelial neoplasia, immature squamous metaplasia, and inflammatory lesions.
- A biopsy, guided by colposcopy, becomes vital when the presence of neoplasia is uncertain.
- Recognizing dense and opaque aceto-white regions, particularly near the squamo-columnar junction, is essential for detecting intraepithelial neoplasia.

Characteristics of CIN (Cervical Intraepithelial Neoplasia):

- Low-grade CIN: Manifests as thin aceto-white lesions with irregular or feathered margins.
- High-grade CIN: These regions are more pronounced—thicker and more opaque with distinct boundaries. Their expansion might reach the endocervical canal, and they exhibit a rough, nodulated texture. Variability in color intensity can be noted within these lesions.

Vascular observations play a pivotal role:

- Both fine and pronounced vascular features, such as punctations and mosaics, are mostly confined to aceto-white areas.
- Low-grade malignancies often show fine punctations or mosaics.

- Conversely, coarse punctations or mosaics hint at high-grade lesions.
- Utilizing green filters can significantly enhance vascular visibility.

For more precise and standardized assessments, incorporating scoring systems like the Swede score [6] can offer valuable guidance in colposcopic evaluations and determinations.

2.1.2 Swede score evaluation

The Swede score [6] is an established metric used in colposcopic evaluations. It provides a systematic approach to assess cervical lesions based on specific characteristics. Each characteristic is scored according to the criteria presented in Table 1, and the cumulative score predicts the severity of the lesion according to the brackets presented in Table 2.

Characteristics	**0**	**1**	**2**
Uptake of acetic acid	Zero or transparent	Shady, milky (not transparent, not opaque)	Distinct, opaque white
Margins and surface	Diffuse	Sharp but irregular, jagged, "geographical". Satellites	Sharp and even; difference in surface level, including "cuffing"
Vessels	Fine, regular	Absent	Coarse or atypical
Lesion size	< 5 mm	$5 - 15$ mm or spanning 2 quadrants	> 15 mm or spanning $3 - 4$ quadrants, or endocervically undefined
Iodine staining	Brown	Faint or patchy yellow	Distinct yellow

Table 1: Swede score assessment

Score	Colposcopic Prediction
0–4	Low-grade/CIN 1
5–6	High-grade/non-invasive cancer/CIN2+
7–10	High-grade/suspected invasive cancer/CIN2+

Table 2: Interpretation of Swede score

2.1.3 Classification of the transformation zone (TZ) in Colposcopy

In colposcopic evaluations, the visibility and positioning of the squamocolumnar junction play a crucial role in categorizing the transformation zone. On the basis of this, the transformation zone can be systematically classified as:

Type 1: The transformation zone, which encompasses the entire squamocolumnar junction, is located in the ectocervix. In simpler terms, the entirety of the upper limit of the TZ is ectocervical.

Type 2: The upper boundary of the TZ is partially or entirely observed within the canal, ensuring visibility throughout a 360-degree angle.

Type 3: The upper boundary of the TZ remains elusive, implying that the upper limit is not visible during examination.

2.1.4 Categorization of aceto-white changes in abnormal colposcopic findings

Post the application of acetic acid during colposcopy, typical aceto-white changes manifest, helping identify potential abnormalities. These can be grouped based on severity as:

Minor (Grade 1): This category predominantly presents with:

- A slender aceto-white epithelium complemented by an irregular, 'geographical' boundary.
- Presence of delicate structures like fine mosaic and fine punctation patterns, indicating lesser severity.

Major (Grade 2): More severe changes in this category are characterized by:

- A pronounced, dense aceto-white epithelial layer that showcases aceto-whitening rapidly upon acid application.

- Noticeable cuffed crypt or gland openings, indicative of potential concerns.
- The epithelium may exhibit coarse mosaic and punctation patterns. In addition, distinct features like a sharp border, the inner border sign, and ridge sign further solidify its classification as major changes.

2.2 Dataset and automatic processing

During our research phase, we sourced 527 colposcopy images from 100 medical records. Expert specialists segmented and annotated each image to differentiate between healthy and pathological cervical tissues.

The segmented and annotated image set, Swede scores and the accompanying code are available at the following address: https://github.com/iclx/AnnoCerv. This work is licensed under a Creative Commons Attribution 4.0 International License [1].

2.2.1 Dataset structure and format description

The organization of the dataset is hierarchical, ensuring ease of navigation and clarity. Here is a detailed breakdown of its structure:

Directory Structure: Each individual case is encapsulated within its own unique folder, named “Case ID”.

Image Files: Within each case folder, there are one or more cervix images saved in the JPG format. The images within a single case can encompass various types, namely acetic acid images, iodine images, and green-filtered images.

Filename convention: The naming convention for the images is standardized for clarity. It consists of a case identifier, followed by the image type, and an index enclosed within parentheses to distinguish multiple images of the same type.

Annotation files: Each acetic acid image (denoted by ‘Aceto’ in the filename) has a corresponding PNG annotation file. This file carries the same primary filename but with a .png extension. For instance, an image named C1Aceto (1).jpg has its annotations in C1Aceto (1).png.

Annotation encoding: The PNG annotation files utilize a specific color encoding to represent various observed features:

- **blue** for the squamous-cylindrical junction,

[1] http://creativecommons.org/licenses/by/4.0/

- **purple** for aceto-white areas,
- **red** for atypical vessels and punctations,
- **brown** for mosaics,
- **yellow** for Naboth cysts, and
- **black** for cuffed gland openings.

The background of these PNG files is transparent. In scenarios where the medical professional did not detect any notable features, the PNG remains entirely transparent without any colored pixels.

Swede scores are cataloged in the CSV file “swede_scores.csv”, where each row corresponds to the score of its respective case.

2.2.2 Automated processing

Given the clearly delineated Dataset Structure and Format Description, the systematic processing of the image set becomes inherently straightforward from a computational perspective. The procedure entails the following methodical steps:

Directory iteration: We commence by traversing each folder, wherein every individual folder signifies a distinct case, warranting content exploration and analysis.

Image type verification: In each case folder, we check for the presence of image types that are of interest for our analysis.

Annotation examination: For every Aceto image, we open its associated PNG file. This step helps us identify different pixel colors, which correspond to specific medical notes.

Statistical aggregation: After collecting all the required data, we can proceed to calculate statistics of interest.

In code listing 1, we exemplify this process to determine several pertinent statistics:

1. The number of cases containing iodine images.
2. The number of cases containing green-filtered images.
3. The number of cases where the squamous-cylindrical junction is not visible (evidenced by the absence of blue pixels).

4. The total count of cases exhibiting atypical regions (atypical vessels, Naboth cysts or cuffed gland openings).

Source code 1: Case processing and statistics extraction – Python code snippet

```python
# Iterate through each folder (case)
for folder in os.listdir(base_path):
    folder_path = os.path.join(base_path, folder)

    print(f'Processing folder {folder}')

    if os.path.isdir(folder_path):
        iodine_present = False
        green_present = False
        blue_absent = True
        atypical_regions = 0

        # Check for image types and corresponding annotations
        for file in os.listdir(folder_path):
            print(f'\tProcessing {file}')
            if "Iod" in file and file.endswith(".jpg"):
                iodine_present = True
            elif "Green" in file and file.endswith(".jpg"):
                green_present = True
            elif "Aceto" in file and file.endswith(".jpg"):
                annotation_file = os.path.join(folder_path,
                ↪  file.replace(".jpg", ".png"))

                if os.path.exists(annotation_file):
                    img = Image.open(annotation_file)
                    pixels = list(img.getdata())

                    for pixel in pixels:
                        # Check for transparency (Alpha channel)
                        if len(pixel) == 4 and pixel[3] > 0:
                            if pixel[:3] == colors['blue']:
                                blue_absent = False
                            elif pixel[:3] != colors['purple']:
                                atypical_regions += 1
```

In our effort to promote accessibility, the code is readily available in the GitHub repository as a Google Colab Notebook named "data_summary.ipynb". The notebook can be easily extended or modified to compute different statistics of interest.

2.2.3 Feature extraction and classification

The GitHub repository additionally contains a Google Colab Notebook named “data_modelling.ipynb” that exemplifies foundational operations, serving as a primer for individuals unfamiliar with image processing and machine learning tasks in the domain of medical imaging.

Source code 2: Feature computation – Python code snippet

```python
from skimage import feature, color

def extract_features(img_path):
    img = cv2.imread(img_path)
    intensity = cv2.cvtColor(img, cv2.COLOR_BGR2GRAY)

    # Texture feature using Local Binary Pattern
    texture_r5 = feature.local_binary_pattern(intensity, P=8*5, R=5,
    ↪  method='uniform')

    # Gradient features using Sobel operator
    grad_x = cv2.Sobel(intensity, cv2.CV_64F, 1, 0, ksize=3)
    grad_y = cv2.Sobel(intensity, cv2.CV_64F, 0, 1, ksize=3)

    # Spatial features: simply the x and y coordinates
    x = np.arange(img.shape[1])
    y = np.arange(img.shape[0])
    x, y = np.meshgrid(x, y)

    # Color-based features
    hsv_img = color.rgb2hsv(img)
    hue = hsv_img[:, :, 0]
    saturation = hsv_img[:, :, 1]
    value = hsv_img[:, :, 2]

    # Stack all features together
    features = np.dstack((intensity, texture_r5, grad_x, grad_y, x, y, hue,
    ↪  saturation, value))

    return features
```

The operations demonstrated include:

Dataset download: Copy of the images to the local machine.

Image resizing: A programmatic approach to altering image and annotation dimensions.

Feature extraction : Focused on the classification of pixels representing the squamous-cylindrical junction (depicted as blue pixels in the annotated PNG files). The features extracted encompass: i) Intensity Features: the direct utilization of pixel intensity; ii) Texture Features: just one Local Binary Patterns (LBP) of radius 5 is used; iii) Gradient Features: the magnitude and direction of image gradients are computed; iv) Spatial Features: the pixel's x and y coordinates is stored; v) Color-based Features: the extraction of Hue, Saturation, and Value (HSV) from pixels. The operations are also depicted in the Code Listings 2 and 3.

Feature scaling: a common operation before fitting the models to the data. By normalizing the features to a consistent scale, we ensure that each one contributes appropriately to the model's outcomes, facilitating faster convergence for gradient-based methods and potentially boosting the model's overall performance.

Data preparation: Given the imbalanced nature of the classification task, the notebook exemplifies a basic balancing method via undersampling. Also, the dataset is divided into training and testing subsets.

Feature correlation analysis and pair plot technique: These methods help assesing the relationships between different features and the target classification variable. Feature correlation analysis quantifies the interdependencies, offering insights into the intrinsic structure of the data. The pair plot technique visualizes pairwise relationships in a dataset. Together, these tools facilitate a comprehensive understanding of the data, and can guide the selection and prioritization of the most pertinent features for model training.

Model training: A Random Forest classifier, utilizing its default parameters, is trained on the extracted, balanced small dataset.

Evaluation: The computation of relevant metrics such as the F1 score and the ROC curve is exemplified.

It is important to emphasize that the methods and techniques highlighted in the notebook are foundational, designed primarily to serve as a rapid, cloud-based experimentation tool for newcomers and enthusiasts. While they offer a convenient starting point for those new to the field, they do not embody the cutting-edge of current research or advanced methodologies. The primary aim is to demonstrate a workflow that is accessible and can be tried out and executed in the cloud in a matter of minutes.

Source code 3: Feature extraction for each case – Python code snippet

```
# Collect data and labels
X_data = []
y_labels = []

for img_file in os.listdir(output_folder):
    if img_file.endswith('.jpg'):
        print(f'\tComputing features for {img_file}')
        features = extract_features(os.path.join(output_folder, img_file))
        X_data.append(features)

        annotation_path = os.path.join(output_folder, img_file.replace('.jpg',
        ↪  '.png'))
        img = cv2.imread(annotation_path)
        annotation = np.array(img[:, :, 0] == 255) & np.array(img[:, :, 2] ==
        ↪  0)
        # Convert blue pixels to label 1, others to 0
        is_junction = annotation.astype(np.int)
        y_labels.append(is_junction)

X_data = np.array(X_data).reshape(-1, 9)
y_labels = np.array(y_labels).reshape(-1)
```

3 Results and discussion

3.1 Segmented and annotated images

To provide insight into our database, we display representative examples of segmented images in Figures 1, 2, 3, and 4, with the annotations superimposed on the cervix images. These images underscore the variety and depth of the content within the dataset.

Figure 1 showcases squamous-cylindrical junctions, aceto white areas, Naboth cyst, punctuation, mosaic, and fine vessels.

In Figure 2, the emphasis is on highlighting the squamous-cylindrical junctions, aceto white areas, polyps, Naboth cysts, and glandular openings.

The rationale for the iodine test is rooted in cellular chemistry: mature squamous epithelium, both original and newly formed, contains glycogen. In contrast, neoplastic and invasive cancer cells typically have minimal or no glycogen. As a result, they do not absorb iodine, appearing as distinct mustard yellow or saffron-colored regions. Following this principle, neoplastic aceto

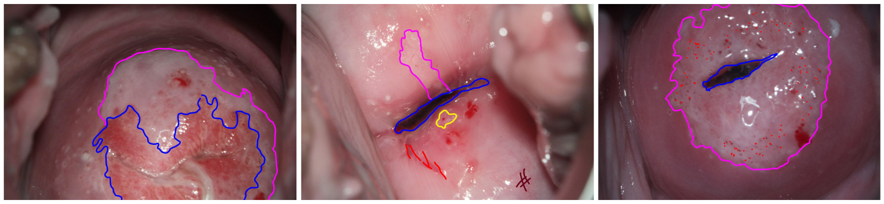

Figure 1: Annotations: blue – squamous-cylindrical junction, purple – aceto white area, red – atypical vessels, punctations, brown – mosaic, yellow – Naboth cyst

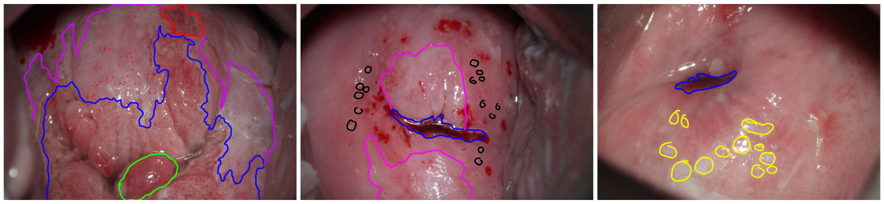

Figure 2: Annotations: blue – squamous-cylindrical junction, purple – aceto white area, yellow – Naboth cysts, black – cuffed gland opening, green – polyp.

white areas remain unaffected by iodine. This characteristic can be observed in Figure 3, where iodine images act as confirmatory markers for suspected lesions.

Colposcopy with a green filter allows visualization of vascular changes. Figure 4 offers insight into this, depicting key vascular alterations like punctuation, mosaic, atypical fine vessels, and larger vessels.

3.2 Exploratory data analysis

Derived from the previously mentioned Google Colab Notebook for data processing, this section delves into patterns and insights within the dataset related to cervical health diagnostics.

The case based image type distribution is presented in Figure 5. A predominant 94% of the cases contain iodine images, highlighting their important role in confirmation and diagnostics. Conversely, green-filtered images, which primarily aid in the evaluation of vascular changes, are present in only 11% of cases. This differential suggests that such vascular evaluations might be less frequently necessitated in the overall diagnostic spectrum. In 13 cases, the squamous-cylindrical junction is not visible, marked by an absence of blue

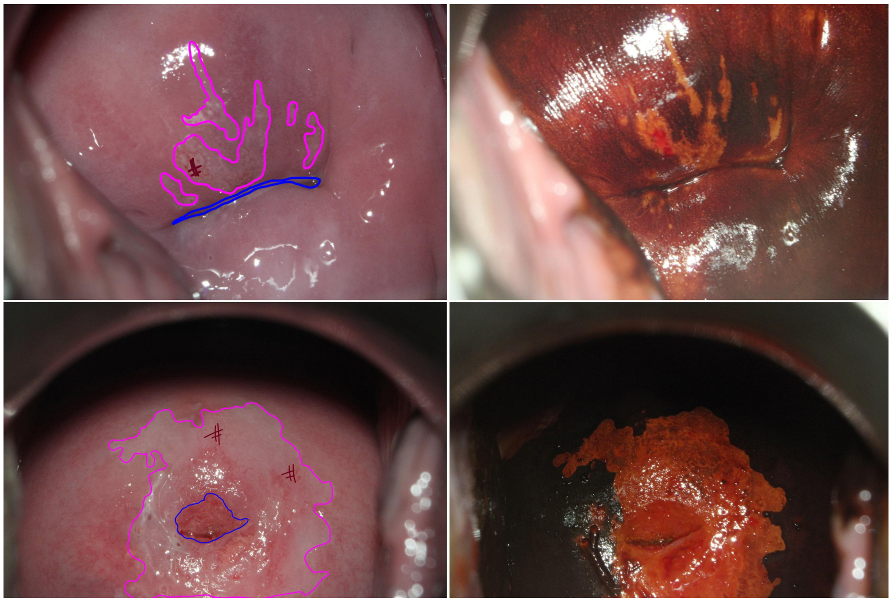

Figure 3: Aceto white neoplastic areas (purple) confirmed with Iodine solution.

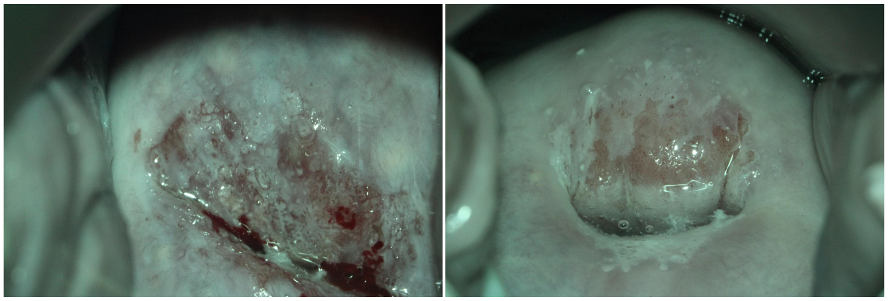

Figure 4: Colposcopy images taken with green filter to highlight vascular changes.

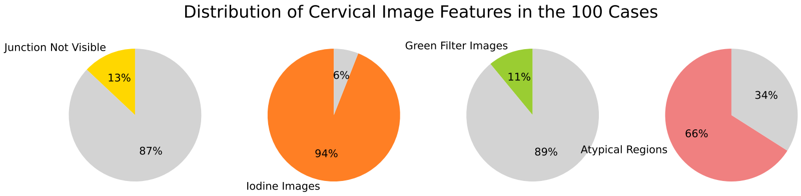

Figure 5: Extracted properties from the 100 colposcopy imaging cases.

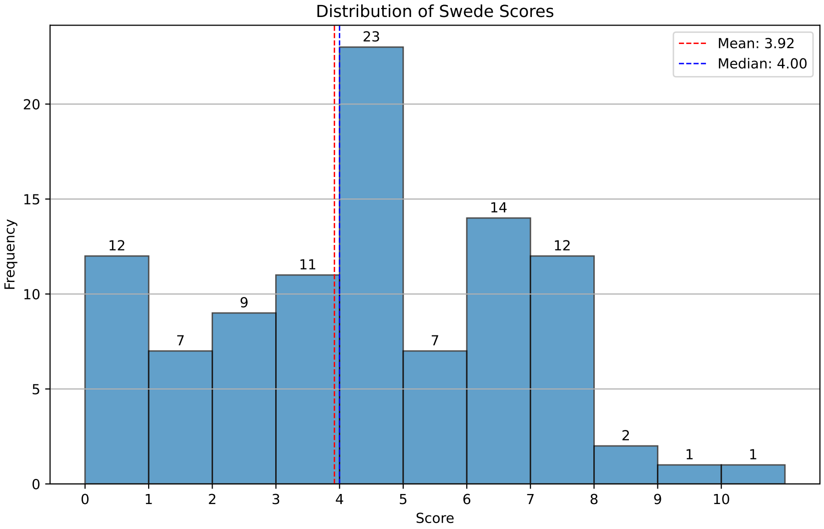

Figure 6: Swede scores distribution and central tendency.

pixels in the annotations. A significant 66% of cases manifest atypical regions, encompassing atypical vessels, Naboth cysts, or cuffed gland openings, underscoring the critical nature of in-depth cervical health assessments.

Figure 6 provides a visual representation of the distribution of Swede scores. Scores, which range from 0 to 10, reveal a spectrum of health conditions. Although 12 cases boast an optimal score of 0, a considerable portion, specifically 23 cases, cluster around a score of 4. However, a handful of cases with high scores of 9 and 10 highlight the existence of severe abnormalities. Statistically, with a mean of 3.92, a median at 4.00, and a standard deviation of 2.40, it is evident that the majority of the cases hover around a moderate risk range.

The dataset offers a detailed snapshot of cervical health through its various image types and score distributions. With atypical regions evident in 66% of cases, the need for meticulous diagnostics becomes even more evident. Although a considerable portion of cases fall within the low-to-moderate risk categories, the presence of high-risk outliers emphasizes the dataset's potential as a valuable resource for training advanced machine learning models. The mix of iodine and green-filtered images within the dataset lays a foundation for exploring a variety of diagnostic methodologies.

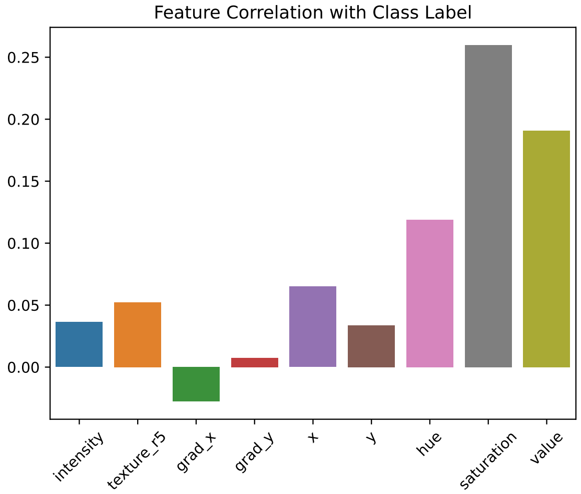

Figure 7: Correlations between the features and the class label.

3.3 Feature correlation

Understanding the relationships between features and the target classification variable is pivotal for effective model building. The two key techniques exemplified in this endeavor are Feature Correlation Analysis and the Pair Plot Technique. The former provides a quantitative measure of interdependencies between features, shedding light on their internal structure and importance. Meanwhile, the Pair Plot Technique offers a visual representation of pairwise relationships across the dataset, enabling a holistic grasp of data intricacies.

A visual representation of the correlations of features with the class label is illustrated in Figure 7. Positional features (mainly 'x') and color-based features stand out with relatively higher positive correlations to the target, suggesting they might play an essential role in classification. In contrast, the feature 'grad_y' shows no correlation, and 'grad_x' seems to be inversely correlated.

Moving onto the Pair Plot in Figure 8, certain observations emerge:

Intensity vs. value: A direct correlation is evident between Intensity (as derived from the gray-scale image) and Value (the brightness of the color). This relationship, expected given that changing the Value in HSV, we are generally increasing or decreasing the brightness of the RGB channels, which in turn will affect the grayscale intensity.

y vs. saturation: A distinctive pattern emerges. Most of the 'Junction' class instances cluster towards the right, indicating a potential joint link between these features and the target variable.

x & y distribution: The scatter plot between x and y showcases a stark disparity in the distribution of our two classes. A potential interpretation is the prevalence of the 'Junction' predominantly towards the center of the image.

Intensity vs. value distribution: Their distribution peaks also differ noticeably, reinforcing the idea of some inherent structural differences within the dataset.

3.4 Classification performance

The Random Forest classifier, using default parameters, served as a baseline to distinguish between 'Non-Junction' and 'Junction' classes in a balanced dataset. Detailed performance metrics are presented in Table 3.

The model achieved an accuracy of around 80%, illustrating a consistent prediction rate for both classes. Precision, which represents the fraction of correct positive predictions, was similar for both classes. The slightly higher recall for the 'Junction' class suggests the model's marginally better ability to detect these instances. With F1-Scores of 0.79 and 0.80 for 'Non-Junction' and 'Junction' respectively, the model demonstrated a balanced performance for both classes, harmonizing precision and recall.

While the current results provide valuable insights, it's worth noting that the model's performance might vary with different configurations or when applied to other datasets. Exploring alternative machine learning algorithms and fine-tuning parameters can potentially unearth more robust classification strategies.

In tandem with the table, Figure 9 visualizes the Receiver Operating Characteristic (ROC) Curve, offering an in-depth view of the performance of the classifier. The curve's area of 0.73 indicates its acceptable discriminative capability, with ample room for improvement. A prominent inflection point at a True Positive Rate (TPR) of 0.8 and a False Positive Rate (FPR) of about 0.37 suggests an optimal threshold. While the TPR is commendable, an FPR of 0.37 highlights the misclassification of a substantial number of negative instances.

The Random Forest classifier, even in its default configuration, yields satisfactory results. Understanding the feature importance provided by the Ran-

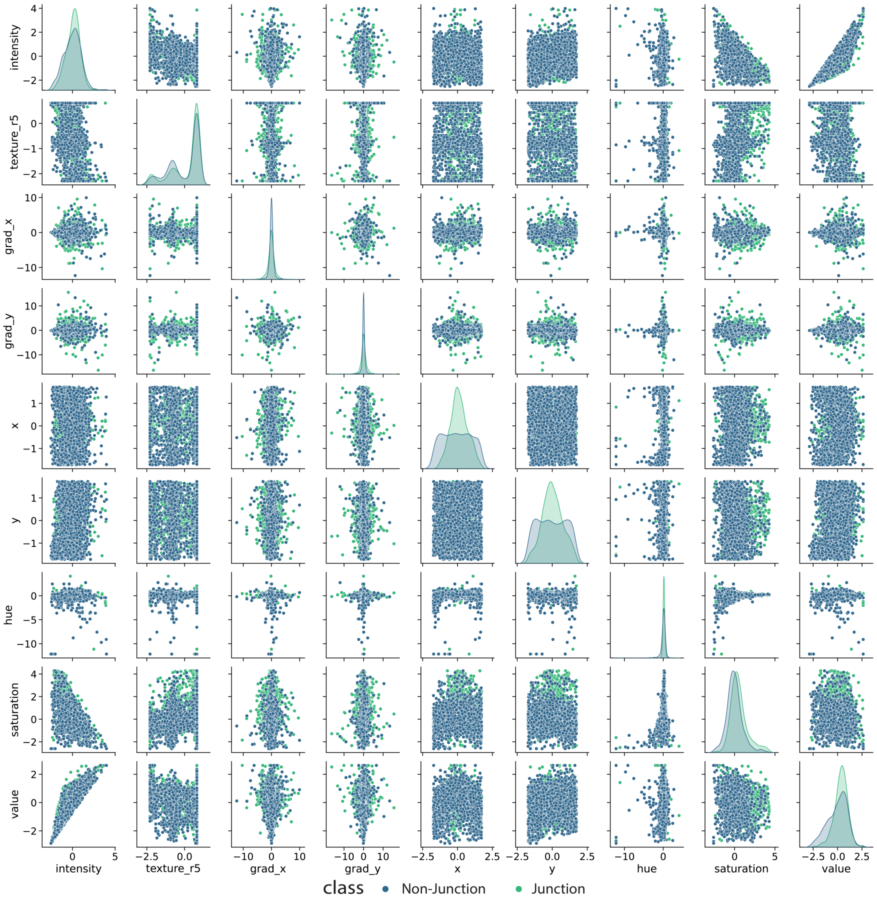

Figure 8: Pairwise relationship of the features.

dom Forest classifier can also offer insights into which variables have a greater influence on the classification decision. A thorough examination of these importance metrics could guide feature engineering efforts, possibly leading to enhanced performance by emphasizing on the most influential features. This introspective approach not only strengthens the model's predictive power but also adds an interpretative dimension to the model, bridging the gap between machine learning predictions and domain-specific knowledge. This study balanced the dataset through subsampling to simplify the classification task for demonstration purposes. For the genuine, heavily imbalanced dataset, har-

	Precision	Recall	F1-Score	Support
Non-Junction	0.80	0.78	0.79	400
Junction	0.79	0.81	0.80	400
accuracy			0.80	800
macro avg	0.80	0.80	0.79	800
weighted avg	0.80	0.80	0.79	800

Table 3: Performance metrics

nessing advanced techniques such as Convolutional Neural Networks (CNNs) and transfer learning could yield superior outcomes.

4 Conclusions

Cervical cancer remains a pressing health concern for women worldwide. While computational methods offer promising avenues for improved diagnosis, their effectiveness is intrinsically linked to the quality and comprehensiveness of available training datasets. Recognizing a discernible gap in this area, we introduce AnnoCerv, a dataset that provides a detailed perspective on cervical colposcopy images. These 527 samples, derived from 100 medical records, present an array of expert-annotated, feature-rich images that aim to support a range of analysis, from basic lesion recognition to Swede score predictions.

AnnoCerv represents our effort to enhance the resources available to researchers and practitioners in the field. While the accompanying code provides an introduction to image processing and machine learning tasks, it's primarily designed for those less familiar with the domain. We acknowledge its foundational nature, emphasizing that there remains a significant opportunity and need for the development of more sophisticated and nuanced methods.

Choosing to present examples via Google Colab Notebooks was a deliberate strategy to enhance accessibility. This approach streamlines the initial setup, allowing users to rapidly interact with the dataset.

We hope that the AnnoCerv image set and code can serve as valuable resources for further research, innovation, and developments in the field of cervical health and diagnostics.

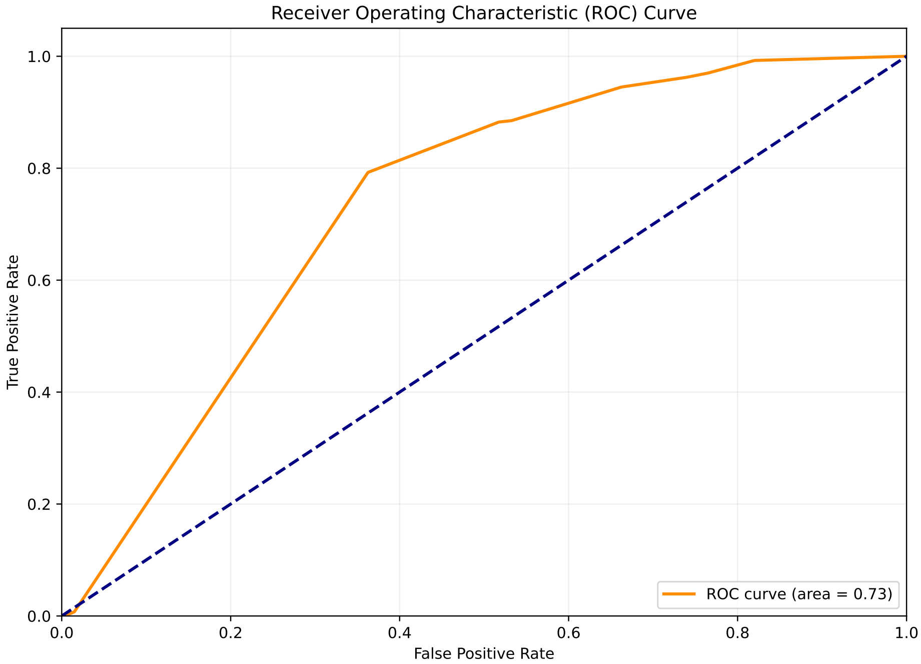

Figure 9: Operating characteristic curve.

Acknowledgements

The work was supported by Financing contract no. 75/221_ap2/21.08.2020, in the Competitiveness Operational Program 2014–2020, Priority Axis 2 - “Information and communication technology (TIC) for a competitive digital economy”, SMIS Code 130106.

References

[1] H.-G. Acosta-Mesa, N. Cruz-Ramírez, R. Hernández-Jiménez, Aceto-white temporal pattern classification using k-nn to identify precancerous cervical lesion in colposcopic images. *Computers in biology and medicine* **39,** 9 (2009) 778–784. ⇒308

[2] O. F. Ahmad, A. S. Soares, E. Mazomenos, P. Brandao, R. Vega, E. Seward, D. Stoyanov, M. Chand, M., L. B. Lovat, Artificial intelligence and computer-aided diagnosis in colonoscopy: current evidence and future directions. *The lancet Gastroenterology & hepatology* **4,** 1 (2019) 71–80. ⇒308

[3] M. Arbyn, F. Verdoodt, P. J. Snijders, V. M. Verhoef, E. Suonio, L. Dillner, S. Minozzi, C. Bellisario, R. Banzi, F.-H. Zhao, et al. Accuracy of human papillomavirus testing on self-collected versus clinician-collected samples: a meta-analysis. *The lancet oncology* **15,** 2 (2014) 172–183. ⇒308

[4] M. N. Asiedu, A. Simhal, U. Chaudhary, J. L. Mueller, C. T. Lam, J. W. Schmitt, G. Venegas, G. Sapiro, G., N. Ramanujam, Development of algorithms for automated detection of cervical pre-cancers with a low-cost, point-of-care, pocket colposcope. *IEEE Transactions on Biomedical Engineering* **66,** 8 (2018) 2306–2318. ⇒308

[5] B. Bai, P.-Z. Liu, Y.-Z. Du, Y.-M. Luo, Automatic segmentation of cervical region in colposcopic images using k-means. *Australasian physical & engineering sciences in medicine*, **41** (2018) 1077–1085. ⇒308

[6] J. Bowring, B. Strander, M. Young, H. Evans,P. Walker, The swede score: evaluation of a scoring system designed to improve the predictive value of colposcopy. *Journal of lower genital tract disease* **14,** 4 (2010) 301–305. ⇒311

[7] B. H. Brown, J. A. Tidy, The diagnostic accuracy of colposcopy–a review of research methodology and impact on the outcomes of quality assurance. *European Journal of Obstetrics & Gynecology and Reproductive Biology* **240** (2019) 182–186. ⇒308

[8] X. Castellsagué, Natural history and epidemiology of hpv infection and cervical cancer. *Gynecologic oncology* **110,** 3 (2008) S4–S7. ⇒308

[9] S. Chilamkurthy, R. Ghosh, S. Tanamala, M. Biviji, N. G. Campeau, V. K. Venugopal, V. Mahajan, P. Rao, P., P. Warier, Deep learning algorithms for detection of critical findings in head ct scans: a retrospective study. *The Lancet* **392,** 10162 (2018) 2388–2396. ⇒308

[10] N. Coudray, P. S. Ocampo, T. Sakellaropoulos, N. Narula, M. Snuderl, D. Fenyö, A. L. Moreira, N. Razavian, A. Tsirigos, Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nature medicine* **24,** 10 (2018) 1559–1567. ⇒308

[11] J. Fan, J. Liu, S. Xie, C. Zhou, Y. Wu, Cervical lesion image enhancement based on conditional entropy generative adversarial network framework. *Methods* **203** (2022) 523–532. ⇒308

[12] A. Goodman Hpv testing as a screen for cervical cancer. *BMJ* **350** (2015). ⇒ 308

[13] P. Guo, Z. Xue, Z. Mtema, K. Yeates, O. Ginsburg, M. Demarco, L. R. Long, M. Schiffman, M., S. Antani, Ensemble deep learning for cervix image selection toward improving reliability in automated cervical precancer screening. *Diagnostics* **10,** 7 (2020) 451. ⇒309

[14] A. Hosny, C. Parmar, J. Quackenbush, L. H. Schwartz, H. J. Aerts, Artificial intelligence in radiology. *Nature Reviews Cancer* **18,** 8 (2018) 500–510. ⇒308

[15] L. Hu, D. Bell, S. Antani, Z. Xue, K. Yu, M. P. Horning, N. Gachuhi, B. Wilson, M. S. Jaiswal, B. Befano, et al. An observational study of deep learning and automated evaluation of cervical images for cancer screening. *JNCI: Journal of the National Cancer Institute* **111,** 9 (2019) 923–932. ⇒308

[16] International Agency for Research on Cancer (IARC). Cervical image bank, 2021. Accessed on 19th October 2023. ⇒309

[17] J. Jin, J. Hpv infection and cancer. *Jama* **319,** 10 (2018) 1058–1058. ⇒308

[18] D. S. Kermany, M. Goldbaum, W. Cai, C. C. Valentim, H. Liang, S. L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell* **172,** 5 (2018) 1122–1131. ⇒308

[19] S. Pimple, G. Mishra, G. Cancer cervix: Epidemiology and disease burden. *Cytojournal* **19** (2022). ⇒307

[20] M. E. Plissiti, P. Dimitrakopoulos, G. Sfikas, C. Nikou, O. Krikoni, A. V. Charchanti, Sipakmed: A new dataset for feature and image based classification of normal and pathological cervical cells in pap smear images. *2018 25th IEEE International Conference on Image Processing (ICIP)* (2018) 3144–3148. ⇒309

[21] W. Prendiville, R. Sankaranarayanan, *Colposcopy and treatment of cervical precancer.* International Agency for Research on Cancer, World Health Organization, 2017. ⇒307, 308

[22] M. Sideri, P. Garutti, S. Costa, P. Cristiani, P. Schincaglia, P. Sassoli de Bianchi, C, Naldoni, L. Bucchi, et al. Accuracy of colposcopically directed biopsy: results from an online quality assurance programme for colposcopy in a population-based cervical screening setting in italy. *BioMed Research International 2015* (2015). ⇒308

[23] M. Underwood, M. Arbyn, W. Parry-Smith, S. De Bellis-Ayres, R. Todd, C. Redman, E. Moss, E. Accuracy of colposcopy-directed punch biopsies: a systematic review and meta-analysis. *BJOG: An International Journal of Obstetrics & Gynaecology* **119,** 11 (2012) 1293–1301. ⇒308

[24] J. Valls, A. Baena, G. Venegas, M. Celis, M. González, C. Sosa, J. L. Santin, M. Ortega, A. Soilán, E. Turcios, et al. Performance of standardised colposcopy to detect cervical precancer and cancer for triage of women testing positive for human papillomavirus: results from the estampa multicentric screening study. *The Lancet Global Health* **11,** 3 (2023) e350–e360. ⇒307, 308

[25] J. M. Walboomers, M. V. Jacobs, M. M. Manos, F. X. Bosch, J. A. Kummer, K. V. Shah, P. J. Snijders, J. Peto, C. J. Meijer, N. Muñoz, Human papillomavirus is a necessary cause of invasive cervical cancer worldwide. *The Journal of pathology* **189,** 1 (1999) 12–19. ⇒308

[26] J. Wang, Analysis of the application values of different combination schemes of liquid-based cytology and high-risk human papilloma virus test in the screening of high-grade cervical lesions. *Brazilian Journal of Medical and Biological Research* **52** (2018). ⇒308

[27] World Health Organization. Cervical cancer – fact sheet, Year. Accessed on 19th October 2023. ⇒307

[28] P. Xue, M. T. A. Ng, Y. Qiao, The challenges of colposcopy for cervical cancer screening in lmics and solutions by artificial intelligence. *BMC medicine* **18** (2020) 1–7. ⇒307, 308

[29] X. Yang, Z. Zeng, S. G. Teo, L. Wang, V. Chandrasekhar, S. Hoi, Deep learning for practical image recognition: Case study on kaggle competitions. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining* (2018), pp. 923–931. ⇒309

[30] Y. Yu, J. Ma, W. Zhao, Z. Li, S. Ding, S. MSCI: A multistate dataset for colposcopy image classification of cervical cancer screening. *International journal of medical informatics* **146** (2021) 104352. ⇒309

Received: August 18, 2023 • *Revised: October 21, 2023*