

# Shape-based Features Investigation for Preneoplastic Lesions on Cervical Cancer Diagnosis

Daniela C. Terra<sup>1,4</sup><sup>a</sup>, Adriano C. Lisboa<sup>2</sup><sup>b</sup>, Mariana T. Rezende<sup>3</sup><sup>c</sup>, Claudia M. Carneiro<sup>3</sup><sup>d</sup>  
and Andrea G. C. Bianchi<sup>4</sup><sup>e</sup>

<sup>1</sup>Department of Computing, Federal Institute of Minas Gerais, Ouro Branco, MG, Brazil

<sup>2</sup>Research Department, GAIA, Belo Horizonte, MG, Brazil

<sup>3</sup>Clinical Analysis Department, Federal University of Ouro Preto, Ouro Preto, MG, Brazil

<sup>4</sup>Department of Computing, Federal University of Ouro Preto, Ouro Preto, MG, Brazil

**Keywords:** Cervical Cancer, Image Classification, Morphological Features, Features Selection, XGBoost Classifier.

**Abstract:** The diagnosis of cervical lesions is an interpretative process carried out by specialists based on cellular information from the nucleus and cytoplasm. Some authors have used cell nucleus detection and segmentation algorithms to support the computer-assisted diagnosis process. These approaches are based on the assumption that the nucleus contains the most important information for lesion detection. This work investigates the influence of morphological information from the nucleus, cytoplasm, and both on cervical cell diagnosis. Experiments were performed to analyze 3,233 real cells extracting from each one 200 attributes related to size, shape, and edge contours. Results showed that morphological attributes could accurately represent lesions in binary and ternary classifications. However, identifying specific cell anomalies like Bethesda System classes requires adding new attributes such as texture.

## 1 INTRODUCTION

Cervical cancer is the fourth most common cancer in women after breast, colorectal, and lung cancer. In 2018 about 570,000 women were diagnosed with the cervical disease, and 311,000 of them died due to malignancy tumors worldwide (Das, 2021). This occurs even with slow progress from precursor lesions to the cancer stage. Thus, the cure of malignancy cases depends on a timely diagnosis or screening for pre-neoplastic lesions. If detected early, the prognosis can be substantially improved with effective treatment (Williams, 2021).

A Pap smear is a cost-effective technique widely used to prevent cervical cancer. Under the microscope, professionals identifying suspicious cell structures following diagnosis protocols internationally adopted such as the Bethesda System (Nayar and Wilbur, 2015). The main disadvantage of such man-

ual analysis is the high rate of false negatives. Screening and diagnosis are subject to misinterpretation by visual habituation and a need for expertise.

Computer-aided diagnostics can reduce errors and increase productivity in cancer screening. Proposals for automated cytology include solutions to detect (Diniz et al., 2021c; Li et al., 2021), segment (Umadi et al., 2020; Teixeira et al., 2022; Zhao et al., 2022), and automate the screening of cell lesions.

Automatic cervical lesion classification follows cell detection or segmentation. The solutions often employ features extraction related to cell size and shape such as area, perimeter, elongation (major/major axes), circularity, and nucleus-cytoplasm ratios (Jantzen et al., 2005; Marinakis et al., 2009; Chankong et al., 2014; Dong et al., 2020; Yakkundimath et al., 2022). Other works measures such as the fractal dimension (Bhowmik et al., 2018), the relative position of the nucleus within cytoplasm (Mariarputham and Stephen, 2015), roughness index, the standard deviation of radial distance and Fourier descriptors (Zhang et al., 2014). Diniz et al. (2021b) uses the CRIC base (Rezende et al., 2021) and traditional ML techniques to classify pre-neoplastic le-

<sup>a</sup> <https://orcid.org/0000-0002-2828-8275>

<sup>b</sup> <https://orcid.org/0000-0001-5773-2200>

<sup>c</sup> <https://orcid.org/0000-0002-9514-9312>

<sup>d</sup> <https://orcid.org/0000-0002-6002-857X>

<sup>e</sup> <https://orcid.org/0000-0001-7949-1188>

sions using cell nuclei shape and texture features. Classification without segmentation is leveraged by deep learning approaches (Dong et al., 2020; Rahman et al., 2021). The work proposed by Diniz et al. (2021a) achieved a high recall rate on detecting cellular lesions using an ensemble of deep neural networks tested with the CRIC dataset images.

This work proposes an investigation of the influence of morphological cell attributes during the classification of cervical cell lesions. The extracted features are related to the size, shape, and edge contours calculated for the nucleus, the cytoplasm, and both components. Our solution is based on traditional machine learning (ML) techniques to classify cervical cells with or without (pre)neoplastic lesions. We evaluated a binary classifier (normal/abnormal), a ternary classifier (normal cell/low-grade lesions/high-grade lesions), and an identifier for the 6 (six) classes of the Bethesda System for cytological diagnosis.

To the best of our knowledge, this is the first approach to investigate the adequacy or limitation of these attributes in automated diagnosis. The main contributions of our investigation are:

- Exploring the potential of shape-based features on discrimination of cervical cells lesions;
- Verify the effectiveness of Elliptic Fourier Descriptors (EFD) in this classification process;
- Evaluate the proposed solution on real images of conventional cytology;
- Analyzing the results of the shape-based classification at the level of cells (both nucleus and cytoplasm), only nuclei and cytoplasms.

The next section discusses our proposal in detail. Section 3 presents experiments and results. Finally, Section 4 reviews the proposed solution’s results.

## 2 METHODOLOGY

This section presents the materials and methods considered. Section 2.1 presents the database used in experiments. Section 2.2 describes the extracted features and the feature selection procedure. Section 2.3 explains the computational model built for the experiments.

### 2.1 Dataset

In this work, we use the *CRIC Cervix-Seg* database of conventional cytology (Rezende et al., 2021). The database contains 3,224 segmented cellular nuclei and cytoplasm from 400 real Pap smear images.

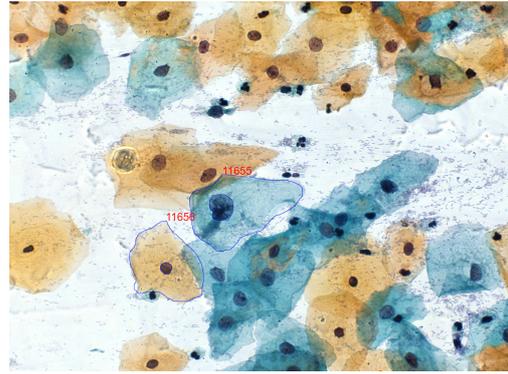


Figure 1: CRIC Cervix-Seg example for nuclei and cytoplasm segmentation.

Classification and segmentation of cells were performed according to Bethesda nomenclature and carried out manually by experienced cytopathologists from the Center for Recognition and Inspection of Cells (CRIC) (see Figure 1).

The Cervix-Seg collection includes six (6) classes: (a) negative for intraepithelial lesion or malignancy (NILM); (b) atypical squamous cells of undetermined significance, possibly non-neoplastic (ASC-US); (c) low-grade squamous intraepithelial lesion (LSIL); (d) atypical squamous cells which cannot exclude high-grade lesions (ASC-H); (e) high-grade squamous intraepithelial lesion (HSIL); and (f) squamous cell carcinoma (SCC).

Table 1 presents the classification groups considered for computational experiments. Our model was built to label cells considering the binary classification (normal and abnormal), the ternary classification (normal cells, low-grade lesions, and high-grade lesions), and the classification based on the Bethesda nomenclature (6 classes).

Table 1: Three classification categories with the number of class samples.

Binary	Ternary	Bethesda	N° of samples
Normal		NILM	862
	Low grade	ASC-US	286
LSIL		536	
Abnormal	High grade	ASC-H	598
		HSIL	874
		SCC	77
Total:			3,233

For the binary categorization, the abnormal cells comprise all Bethesda labels except NILM (normal). Another possible classification is used to group Bethesda categories into 3 (three) classes: normal

cells (NILM), low-grade lesions (ASC-US and LSIL), and high-grade lesions (ASC-H, HSIL, and SCC). Low and high-grade groupings become important due to different treatment protocols. For low-grade lesions, the follow-up requires a repeat screening. In the case of high-grade lesions cells, patients should undergo colposcopy and/or biopsy (Sung et al., 2021).

## 2.2 Shape-based Features

We used 200 features for each cell. The same 98 measurements applied to the nucleus were calculated for the cytoplasm. The remaining 4 comprise the two cellular components. Features are related to the size, shape, and edge contour of the nucleus (N) and the cytoplasm (C), and some ratios between N and C:

- Size: area, bounding box, convex hull, perimeter, equivalent diameter (circumference), minor and major axis;
- Shape: circularity, compacity, eccentricity, convexity, solidity, elongation, fractal dimension;
- Contour: roughness index, entropy, kurtosis, and other statistics of normalized radial distance (from the centroid to edge points). Also, the first 20 coefficients of the elliptic Fourier series (Kuhl and Giardina, 1982);
- N/C relations: nucleus relative position (within the cell), nucleus to cytoplasm ratios for the area, perimeter, bounding box, and convex hull.

The Box Counting method was used to calculate the fractal dimension (FD) for the cell components (N and C) (Konatar et al., 2020). As known, the more irregular the regions, the higher the FD value.

Elliptic Fourier coefficients are also related to edge contour irregularities in the frequency spectrum. The EFD method is based on the string code (connectivity 8) extracted from contour points of a region (Kuhl and Giardina, 1982). We use the first 20 EFD coefficients for later feature selection.

Roughness index and standard deviation of radial distance were used for cervical cells by Zhang et al. (2014). We calculated these features as described by Po-Hsiang Tsui et al. (2010), as they are commonly used in breast tumor detection.

### 2.2.1 Feature Selection

To improve performance and get some intuition about the interpretability of the model, we select the most relevant features for each estimator shown in Figure 2.

At most, 30 attributes were chosen from 200 using two methods: mutual information (MI) and simulta-

neous perturbation stochastic approximation (SPFSR) (Akman et al., 2023).

The MI is a filter method based on the statistical measure related to the joint entropy of the variables. The SPFSR is based on stochastic simultaneous perturbation approximation. The SPFSR as a wrapper-based proposal can be used with any classifier or regression to optimize a suitable performance metric. It is a multivariate approach that considers the interactions between features so that redundancies decrease their scores.

## 2.3 Computational Model

Figure 2 presents the proposed model. The procedure starts from the two .csv files. One file contains labels for each cell/image along with the set of contour points  $(x,y)$  of the nucleus (Figure 2,  $f1$ ). The other file is analogous to that of the nucleus for the cell's cytoplasm (Figure 2,  $f2$ ). These points are the contour of the manual segmentation made by cytopathologists and used here to reconstruct the cells' masks for the nucleus and cytoplasm. From masks, all described features are calculated (Figure 2,  $fx$ ).

Before classification, a selection of the most important features was considered for each model estimator (Figure 2,  $fs$ ). The filtered set of features is used as input to the model.

Cells were classified according to Table 1 in two ways: with independent classifiers for 2, 3, and 6 classes (Figure 2, by estimators  $a1$ ,  $a2$ , and  $a3$ , respectively) and with a hierarchical classifier (Figure 2, by estimators  $b1$ ,  $b2$ ,  $b3.1$ , and  $b3.2$ ). Our solution implements a hierarchical classification as proposed by Diniz et al. (2021c).

A hierarchical categorization operates in levels. The first level defines a binary classifier to distinguish normal and abnormal cells (Figure 2,  $b1$ ). Cells identified as abnormal are reclassified as low- or high-grade lesions at the second level (Figure 2,  $b2$ ). Finally, two third-level classifiers must identify specific lesions according to Bethesda nomenclature (Figure 2,  $b3.1$  and  $b3.2$ ). In  $b3.1$ , low-grade lesions will be differentiated as ASC-US or LSIL. In estimator  $b3.2$ , cells with more severe anomalies will be categorized into ASC-H, HSIL, or SCC.

The class predictions of both approaches are evaluated in the end through evaluation metrics (Figure 2,  $ev$ ). We employ commonly used metrics for classification as mentioned by Jiang et al. (2022): accuracy (Acc), precision (P), recall (R), specificity (Spec), and F1-score (F1).

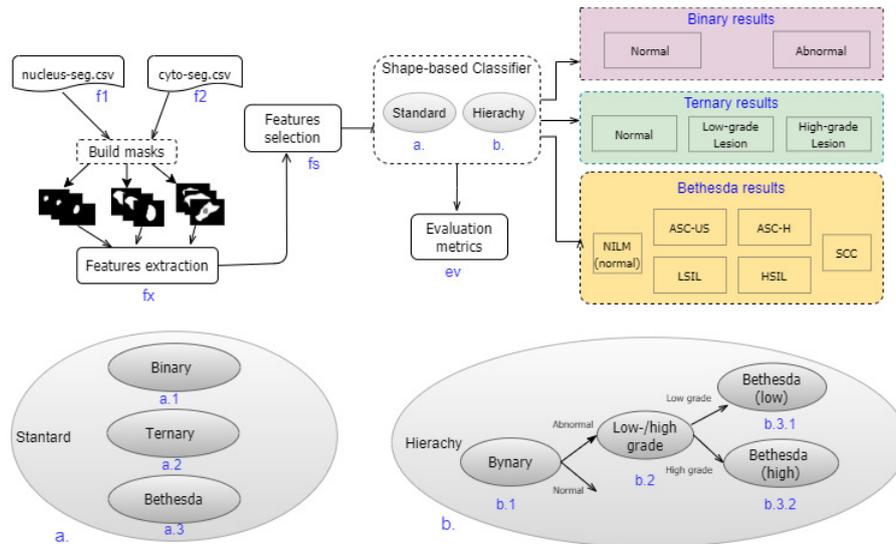


Figure 2: Shape-based diagnostic solution: CRIC-Seg files (f1, f2), attribute extraction (fx) and selection (fs), standard classification (estimators a.1, a.2 and a.3) and hierarchical classification (b1, b2, b3.1 and b3.2 estimators).

### 2.3.1 Learning Algorithms

Each classifier component of Figure 2 is a traditional machine learning (ML) algorithm: Support Vector Machine (SVM), Random Forest (RF), and eXtreme Gradient Boosting (XGBoost).

An SVM outputs an ordered sorted map based on the training data using a subset of those points in the decision function (called support vectors). Such points define the best margins to separate samples in classes in an  $n$ -dimensional space (Geron, 2022).

RF is a bagging method based on decision trees (DC). It introduces randomness in selecting subsamples and features from the data to build the trees. Results have a decreasing variation of the error (Geron, 2022).

XGBoost is another ensemble model based on gradient boosting. Predictions are adjusted sequentially after each weak estimator (for example, a shallow DC). The method improves performance, overfitting, and other flexibilities (Chen and Guestrin, 2016).

### 2.3.2 Oversampling

While the random over-sampler technique duplicates some of the original samples, other techniques build 'synthetic' samples based on original examples (Chawla et al., 2004). The Synthetic Minority Over-sampling Technique (SMOTE) and the Borderline-SMOTE are some of these methods (Chawla et al., 2002; Han et al., 2005). They operate in the feature space rather than at the data level (i.e., the image).

In the SMOTE method, oversampling is done by taking an original sample from the minority class and

introducing a new sample considering any of its nearest  $k$  neighbors through interpolation. The difference between the Borderline-SMOTE method and the original SMOTE is that the former restricts the original samples of the minority class. Selected samples in Borderline-SMOTE are those at the borderline between the minority class and the majority classes.

## 3 RESULTS AND DISCUSSIONS

Experiments were written in Python (version 3.9.1) using well-known libraries for ML and data manipulation/visualization, such as scikit-learn, SciPy, and scikit-image. Other modules were employed for specific tasks. For data augmentation techniques (SMOTE and Borderline-SMOTE) we apply the imbalanced-learn module (<https://imbalanced-learn.org/stable/install.html>). The pyEFD package was used to calculate the Elliptic Fourier coefficients (<https://pyefd.readthedocs.io/en/latest/>). The implementation of the SPFSR method used is available at (<https://github.com/akmand/spFSR>). Codes for the experiments described here are available at <https://github.com/danielaterra/shape-based-CervicalCellsClassifier>.

In all model executions, we used 10-fold cross-validation with the data augmentation techniques defined in Section 2.3.2. Data augmentation was applied within each fold of the training data to equal the number of instances with those of the majority class. The results of the experiment show that the results of SMOTE were slightly superior to those of Borderline-

SMOTE. The evaluation metrics presented below are the results of experiments using SMOTE.

### 3.1 Features Selection Procedure

Calling pyEFD method to retrieve the first 20 EFD coefficients results in 40 values: 20 to  $x$  variations and 20 to  $y$ . All EFD coefficients plus the other features described in Section 2.2 were calculated from each mask of the nucleus and cytoplasm (see Figure 2). At most 30 features were selected from the total: 200 (when considering the nucleus and cytoplasm) or 98 for experiments applied to only one of these cellular structures.

As mentioned in Section 2.2.1 two FS techniques were used: MI and the SPFSR.

Table 2: Feature selection methods: binary results using MI and SPFSR.

		Acc	P	R	Espec	F1
MI	SVM	0.94	0.969	0.954	0.91	0.961
	RF	0.94	0.969	0.964	0.91	0.966
	XGB	0.95	0.971	0.968	0.91	0.969
SPFSR	SVM	0.94	0.969	0.950	0.91	0.959
	RF	0.94	0.968	0.963	0.90	0.965
	XGB	0.94	0.967	0.961	0.90	0.964

Figure 3 presents SPFSR features selected for the normal/abnormal classification. The scores suggest the most discriminatory attributes: 1) area of the convex hull and the edges entropy of the nuclei; 2) compacity and area of the cytoplasm (the larger, the less circular or irregular); and (3) cells' nucleus-to-cytoplasm ratio (N/C).

Table 2 shows the classification metrics of the binary prediction using both methods. As the values were similar, hereafter the best prediction is shown.

### 3.2 Experiments

We performed 3 (three) tests with the proposed solutions using a different set of features: 1) shape features selected considering nucleus, cytoplasm, and both (N/C); 2) nucleus shape features (N); and 3) cytoplasm shape features (C). The first tests perform the classification using the standard solution.

#### 3.2.1 Standard Classification

Table 3 presents evaluation metrics for the normal/abnormal classification using attributes for the cell (N/C), the nucleus (N), and the cytoplasm (C). Results for detections of normal cells, low-grade or high-grade lesions are shown in Table 4. Table 5 presents the results for the 6 (six) classes detection.

Table 3: Binary classification: from cells, nuclei, and cytoplasm.

		Acc	P	R	Spec	F1
1. N/C	SVM	0.942	0.971	0.951	0.917	0.960
	RF	0.952	0.970	0.966	0.912	0.967
	XGB	0.954	0.971	0.968	0.917	0.969
2. N	SVM	0.877	0.952	0.880	0.870	0.912
	RF	0.874	0.949	0.878	0.863	0.910
	XGB	0.884	0.937	0.906	0.822	0.919
3. C	SVM	0.901	0.970	0.894	0.922	0.927
	RF	0.835	0.939	0.830	0.849	0.870
	XGB	0.884	0.944	0.897	0.847	0.916

Table 4: Ternary classification: from cells, nuclei, and cytoplasm.

		Acc	P	R	Spec	F1
1. N/C	SVM	0.936	0.939	0.936	0.968	0.935
	RF	0.938	0.941	0.938	0.968	0.936
	XGB	0.941	0.943	0.941	0.970	0.939
2. N	SVM	0.695	0.708	0.695	0.843	0.689
	RF	0.713	0.724	0.713	0.852	0.709
	XGB	0.711	0.720	0.711	0.848	0.707
3. C	SVM	0.915	0.917	0.915	0.957	0.913
	RF	0.846	0.849	0.846	0.926	0.844
	XGB	0.884	0.888	0.884	0.943	0.882

Table 5: Bethesda classification: cells, nuclei, and cytoplasm.

		Acc	P	R	Spec	F1
1. N/C	SVM	0.632	0.693	0.632	0.927	0.640
	RF	0.658	0.700	0.658	0.931	0.663
	XGB	0.682	0.698	0.682	0.935	0.682
2. N	SVM	0.442	0.482	0.442	0.888	0.437
	RF	0.475	0.493	0.475	0.892	0.469
	XGB	0.490	0.489	0.490	0.893	0.479
3. C	SVM	0.620	0.664	0.620	0.924	0.623
	RF	0.557	0.601	0.557	0.910	0.563
	XGB	0.619	0.634	0.619	0.921	0.617

#### 3.2.2 Hierarchical Classification

As observed in Table 5 the model proved to be unfeasible for a Bethesda diagnosis. Trying to improve and analyze the results we implemented the hierarchical solution as depicted in Section 2.3. Tables 6-8 exhibit the results for the 2, 3, and 6 classes, respectively. (Small differences in the results of tables 3 and 6 are due to the use of a 10-stratified fold considering binary and Bethesda labels, respectively.)

Experiments confirmed that the hierarchical solution did not solve class confusion for adequate prediction of specific lesions according to Bethesda System. Figure 4 shows the results of the hierarchical classifier in a confusion matrix. Note that most errors occurred within the low-grade (ASCUS/LSIL) and high-grade (ASCH/HSIL) lesion categories. Likewise, Carcinoma diagnoses were more frequently confused with ASCH and HSIL.

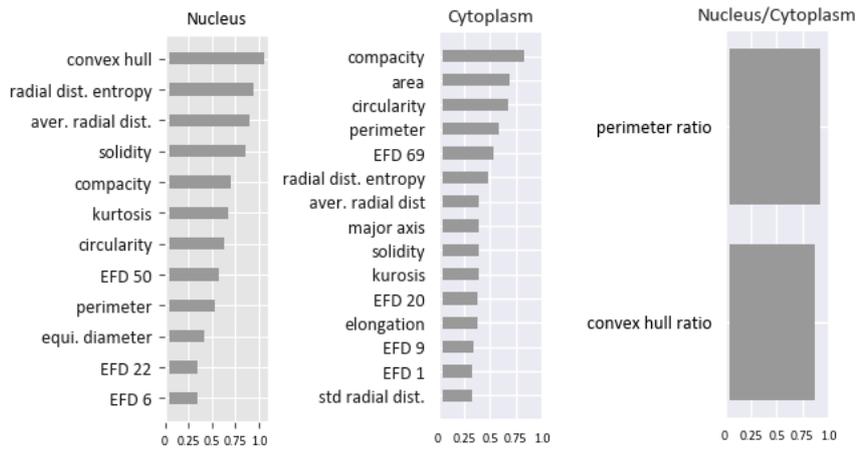


Figure 3: SPFSR relative scores: features selected to normal/abnormal classifier (Figure 2, a.1).

Table 6: Binary (hierarchical) classification: from cells, nuclei, and cytoplasm.

		Acc	P	R	Spec	F1
1. N/C	SVM	0.944	0.969	0.955	0.916	0.962
	RF	0.961	0.968	0.978	0.913	0.973
	XGB	0.960	0.969	0.977	0.915	0.973
2. N	SVM	0.879	0.950	0.881	0.872	0.914
	RF	0.879	0.950	0.881	0.874	0.914
	XGB	0.887	0.933	0.911	0.822	0.922
3. C	SVM	0.9122	0.968	0.910	0.917	0.938
	RF	0.8413	0.933	0.843	0.835	0.886
	XGB	0.8797	0.930	0.903	0.813	0.916

Table 7: Ternary (hierarchical) classification: from cells, nuclei, and cytoplasm.

		Acc	P	R	Spec	F1
1. N/C	SVM	0.928	0.929	0.928	0.964	0.928
	RF	0.945	0.945	0.945	0.971	0.945
	XGB	0.944	0.944	0.944	0.971	0.944
2. N	SVM	0.688	0.695	0.688	0.842	0.683
	RF	0.706	0.712	0.706	0.851	0.702
	XGB	0.701	0.701	0.701	0.843	0.700
3. C	SVM	0.897	0.903	0.897	0.949	0.897
	RF	0.836	0.847	0.836	0.921	0.831
	XGB	0.885	0.886	0.885	0.944	0.885

Table 8: Bethesda (hierarchical) classification: from cells, nuclei, and cytoplasm.

		Acc	P	R	Spec	F1
1. N/C	SVM	0.631	0.665	0.631	0.926	0.642
	RF	0.685	0.698	0.685	0.935	0.689
	XGB	0.688	0.684	0.688	0.935	0.685
2. N	SVM	0.425	0.458	0.425	0.885	0.416
	RF	0.479	0.473	0.479	0.893	0.469
	XGB	0.490	0.479	0.490	0.893	0.483
3. C	SVM	0.617	0.647	0.617	0.923	0.623
	RF	0.593	0.596	0.593	0.915	0.585
	XGB	0.629	0.622	0.629	0.923	0.625

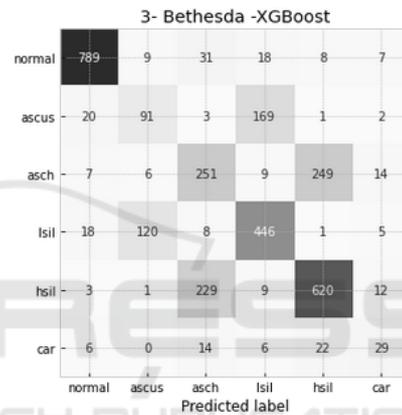


Figure 4: Confusion matrix of a Bethesda classification for N/C features using Random Forest.

### 3.2.3 Fourier Coefficients Results

Table 9 shows the results of running the model using only EFD features as cell shape descriptors. We confirm the irrelevance of these descriptors by again running the solution with all other features except the EFD coefficients. The results are shown in Table 10. As shown in Figure 3, EFD did not well appear to explain cervix lesions as received low scores by feature selectors (e.g., SPFSR).

Table 9: EFD-based classification results: 30 descriptors from cells (binary).

	Acc	P	R	Spec	F1
1. N/C (XGBoost)	0.78	0.86	0.84	0.61	0.85

Table 10: Classification without EFD: from cells (binary).

		Type	Acc	P	R	Spec	F1
1. N/C	binary		0.95	0.97	0.96	0.91	0.96
	Ternary		0.94	0.94	0.94	0.97	0.94

### 3.3 Discussions

As Diniz et al. (2021c) pointed out, the recognized correlation of cervical lesions with toxicological changes in the nucleus allows an analysis based only on this components to classify the degree of lesions. However, the results presented for a classification based on morphology suggest the influence of the cytoplasm in the diagnosis. Furthermore, we observe that:

- Morphology attributes, as proposed in this work, can assist a cytopathologist's final diagnosis for binary (normal/abnormal) and ternary (normal/low-/high grade) classifications. The F1-score values from cell tests (N/C) are above 92% (see Tables 3, 4, 6, and 7);
- The confusion matrix in Figure 4 confirms that most errors fall within subcategories of low/high-grade lesions. Despite the failure, it is worth remembering that the same clinical procedure must be applied in cases of ASC-US and LSIL (low grade) and in cases of ASC-H and HSIL (high grade).
- The nucleus/cytoplasm ratio attributes for the area, perimeter, and convex hull received high scores in the feature selection procedures. These features always contributed to the prediction results for cells (N/C) (see Tables 3, 4, 6 and 7).

## 4 CONCLUSION

This is the first work to validate a classification based only on morphological attributes. A model for classifying cervical cell lesions was evaluated according to Bethesda System's diagnostic classes. We extract 200 features related to the size, shape, and edge contour of each cell from a total of 3,233 samples from a real Papanicolaou image dataset (CRIC Cervix-Seg). As can be seen in Tables 9 and 10 the discrimination test results confirm that Elliptic Fourier Descriptors (EFD) as features showed a result lower than expected.

Table 11 presents our shape-based solution results compared to other works (Diniz et al., 2021b,a) for cervix cell diagnosis. Both solutions employed only cellular nuclei from image patches of the CRIC dataset, and both performed texture analysis.

Our work suggests the dependence of other types of attributes, such as the texture of the nucleus and cytoplasm for discrimination of specific degrees of lesions as the 6 (six) classes considered here from Bethesda nomenclature. However, the proposed morphological attributes play an important role in

Table 11: Comparison with the methods from the literature.

Method	N° classes	Acc	P	R	Spec	F1
1.	2	<b>0.96</b>	<b>0.97</b>	<b>0.97</b>	0.91	<b>0.97</b>
	3	0.94	0.94	0.94	0.97	0.94
	6	0.68	0.68	0.68	0.93	0.68
2.	2	0.95	0.95	0.95	0.95	0.95
	3	0.96	0.96	0.96	0.97	0.96
	6	0.96	0.91	0.90	0.98	0.90
3.	2	0.96	0.96	0.96	<b>0.96</b>	0.96
	3	0.96	0.94	0.94	0.97	0.94
	6	0.95	0.85	0.85	0.97	0.85

1. Our proposal/ XGBoost,RF/ cells' morphology
2. Diniz et al. (2021b)/ RF / nuclei texture/shape
3. Diniz et al. (2021a)/ DNN/ cells' image

binary (normal/abnormal) and ternary (normal/low grade/high grade) classification. As shown in Table 11, the results of our shape-based proposal for 2 (two) and 3 (three) classes were comparable to existing works using the CRIC dataset.

## ACKNOWLEDGEMENTS

Daniela C. Terra acknowledges the support of the Federal Institute of Minas Gerais (IFMG). We also acknowledge Federal University of Ouro Preto (UFOP), FAPEMIG [APQ-00751-19, APQ-01306-22, APQ-01518-21, PPSUS-FAPEMIG/APQ-03740-17]; CNPq [303266/2019-8, 305895/2019-2, 308947/2020-7]; Pró-Reitoria de Pesquisa, Pós-Graduação e Inovação - PROPP/UFOP [19/2020, 23109.000928/2020-33, and 23109.000929/2020-88]; CAPES, and Ministry of Health [905103/2020].

## REFERENCES

- Akman, D. V., Malekipirbazari, M., Yenice, Z. D., Yeo, A., Adhikari, N., Wong, Y. K., Abbasi, B., and Gumus, A. T. (2023). k-best feature selection and ranking via stochastic approximation. *Expert Systems with Applications*, 213:118864.
- Bhowmik, M. K., Roy, S. D., Nath, N., and Datta, A. (2018). Nucleus region segmentation towards cervical cancer screening using AGMC-TU Pap-smear dataset. In *ACM Int. Conf. Proceeding Series*, pages 44–53, New York, New York, USA. ACM Press.
- Chankong, T., Theera-Umpon, N., and Auephanwiriyakul, S. (2014). Automatic cervical cell segmentation and classification in Pap smears. *Computer Methods and Programs in Biomedicine*, 113(2):539–556.
- Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journ. of Art. Intel. Research*, 16:321–357.

- Chawla, N. V., Japkowicz, N., and Kotcz, A. (2004). Editorial: Special issue on learning from imbalanced data sets. *SIGKDD Explor. Newsl.*, 6(1):1–6.
- Chen, T. and Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, pages 785–794, New York, NY, USA. ACM.
- Das, M. (2021). WHO launches strategy to accelerate elimination of cervical cancer. *The Lancet Oncology*, 22(1):20–21.
- Diniz, D. N., Rezende, M. T., Bianchi, A. G., Carneiro, C. M., Luz, E. J., Moreira, G. J., Ushizima, D. M., de Medeiros, F. N., and Souza, M. J. (2021a). A deep learning ensemble method to assist cytopathologists in pap test image classification. *Journal of Imaging*, 7(7).
- Diniz, D. N., Rezende, M. T., Bianchi, A. G. C., Carneiro, C. M., Ushizima, D. M., de Medeiros, F. N. S., and Souza, M. J. F. (2021b). A Hierarchical Feature-Based Methodology to Perform Cervical Cancer Classification. *Applied Sciences*, 11(9):2–19.
- Diniz, D. N., Vitor, R. F., Bianchi, A. G. C., Delabrida, S., Carneiro, C. M., Ushizima, D. M., de Medeiros, F. N. S., and Souza, M. J. F. (2021c). An ensemble method for nuclei detection of overlapping cervical cells. *Expert Systems with Applications*, 185:115642.
- Dong, N., Zhao, L., Wu, C., and Chang, J. (2020). Inception v3 based cervical cell classification combined with artificially extracted features. *Applied Soft Computing*, 93:106311.
- Geron, A. (2022). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. O’Reilly Media, Inc.
- Han, H., Wang, W.-Y., and Mao, B.-H. (2005). Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In *Int. conf. on intelligent computing*, pages 878–887. Springer.
- Jantzen, J., Norup, J., Dounias, G., and Bjerregaard, B. (2005). Pap-smear Benchmark Data For Pattern Classification. *Proc. NiSIS 2005, Albufeira, Portugal*, pages 1–9.
- Jiang, H., Zhou, Y., Lin, Y., Chan, R. C., Liu, J., and Chen, H. (2022). Deep learning for computational cytology: A survey. *Medical Image Analysis*, page 102691.
- Konatar, I., Popovic, T., and Popovic, N. (2020). Box-Counting Method in Python for Fractal Analysis of Biomedical Images. *2020 24th Int. Conf. on Information Technology, IT 2020*, (February).
- Kuhl, F. P. and Giardina, C. R. (1982). Elliptic fourier features of a closed contour. *Computer Graphics and Image Processing*, 18(3):236–258.
- Li, X., Xu, Z., Shen, X., Zhou, Y., Xiao, B., and Li, T.-Q. (2021). Detection of Cervical Cancer Cells in Whole Slide Images Using Deformable and Global Context Aware Faster RCNN-FPN. *Current Oncology*, 28(5):3585–3601.
- Mariarpatham, E. J. and Stephen, A. (2015). Nominated Texture Based Cervical Cancer Classification. *Computational and Mathematical Methods in Medicine*, 2015:1–10.
- Marinakakis, Y., Dounias, G., and Jantzen, J. (2009). Pap smear diagnosis using a hybrid intelligent scheme focusing on genetic algorithm based feature selection and nearest neighbor classification. *Computers in Biology and Medicine*, 39(1):69–78.
- Nayar, R. and Wilbur, D. C. (2015). *The bethesda system for reporting cervical cytology: Definitions, criteria, and explanatory notes*. Springer International Publishing.
- Po-Hsiang Tsui, Yin-Yin Liao, Chien-Cheng Chang, Wen-Hung Kuo, King-Jen Chang, and Chih-Kuang Yeh (2010). Classification of Benign and Malignant Breast Tumors by 2-D Analysis Based on Contour Description and Scatterer Characterization. *IEEE Trans. on Med. Imaging*, 29(2):513–522.
- Rahaman, M. M., Li, C., Yao, Y., Kulwa, F., Wu, X., Li, X., and Wang, Q. (2021). DeepCervix: A deep learning-based framework for the classification of cervical cells using hybrid deep feature fusion techniques. *Computers in Biology and Medicine*, 136:104649.
- Rezende, M. T., Silva, R., Bernardo, F. d. O., Tobias, A. H., Oliveira, P. H., Machado, T. M., Costa, C. S., Medeiros, F. N., Ushizima, D. M., Carneiro, C. M., and Bianchi, A. G. (2021). Cric searchable image database as a public platform for conventional pap smear cytology data. *Scientific Data*, 8(1):151.
- Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., and Bray, F. (2021). Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 71(3):209–249.
- Teixeira, J. B. A., Rezende, M. T., Diniz, D. N., Carneiro, C. M., Luz, E. J. d. S., Souza, M. J. F., Ushizima, D. M., de Medeiros, F. N. S., and Bianchi, A. G. C. (2022). Segmentação automática de núcleos cervicais em imagens de Papanicolau. In *Anais do XXII Simp. Bras. de Computação Aplicada à Saúde*, pages 346–357. Soc. Bras. de Computação.
- Umadi, A., Nagarajan, K., Venkatesha, J. B., Ganesh, A., and George, K. (2020). Automated Segmentation of Overlapping Cells in Cervical Cytology Images Using Deep Learning. In *2020 IEEE 17th India Council Int. Conf. INDICON 2020*, pages 1–7. IEEE.
- Williams, A. (2021). Cervical cancer: what’s new in squamous cell neoplasia. *Diagnostic Histopathology*, 27(12):478–482.
- Yakkundimath, R., Jadhav, V., Anami, B., and Malvade, N. (2022). Co-occurrence histogram based ensemble of classifiers for classification of cervical cancer cells. *Journal of Electronic Science and Technology*, 20(3):100170.
- Zhang, L., Kong, H., Ting Chin, C., Liu, S., Fan, X., Wang, T., and Chen, S. (2014). Automation-assisted cervical cancer screening in manual liquid-based cytology with hematoxylin and eosin staining. *Cytometry Part A*, 85(3):214–230.
- Zhao, Y., Fu, C., Xu, S., Cao, L., and Ma, H.-f. (2022). LFANet: Lightweight feature attention network for abnormal cell segmentation in cervical cytology images. *Computers in Biology and Medicine*, 145:105500.