Analysis of Human Cervical Cell Images from Pap Smears for Classification

CAROLINE FERNANDEZ, TANVI PATEL, AND KRUSHANG PANDYA
New York Institute of Technology

Mentor: Niharika Nath PhD, Professor of Biological and Chemical Sciences,
New York Institute of Technology

Abstract

Cervical cancer is the fourth most commonly occurring cancer in women. It can be prevented by regular screenings to find precancerous cells through the Papanicolaou Smear Test and microscopy which can show cellular abnormalities. However, the manual microscopic screening for the nuclear abnormalities is subjective and prone to error, making automated detection a necessity. This study aims to quantify the nuclear features related to shape characteristics of normal and abnormal cells from pap smear images and examine potential detection of multi classes. Using the ground truth images of normal and abnormal cells we extracted the nuclear shape features that corresponded to the classified cells such as normal and three categories of abnormal: mild, moderate and severe; that is four classes. The dataset of the nuclear shape features were visually plotted as a heat map and bubble plots using the ground truth or known predetermined labelled normal, mild-, moderate- or severe abnormal cells, and also without any such labeling. By clustering, 78 - 89% of the cells were successfully matched with the ground truth. Further, we found that more than 4 classes were obtained. In conclusion, by data visualization techniques we can classify precancerous cells.

Keywords: cervical cancer, pap, k-means, cluster, visualization, image analysis

Introduction

In 2020, there were 604,000 new cases of cervical cancer worldwide (American Cancer Society, 2021). Precancerous lesions detected by microscopic inspection of cells scraped from the cervix and stained are used for prevention of cervical cancer (Arbyn et al., 2020). The Papanicolaou Smear Test and microscopy is a commonly used method to detect the presence of precancerous and cancerous cells (Arbyn et al., 2020). A pathologist analyses the cells manually under the microscope to observe any change in the shape of nucleus or the nucleus to cytoplasm ratio. The cells are identified based on the irregular shape and the enhancement of size of the nucleus. However, the visual screening done via Pap Test is not efficient enough since it can be subjective. The tests are prone to human errors, yield false negatives and positives and can take longer durations of times. Also, it takes about 5-10 minutes to analyze each slide and it has been suggested that a cytopathologist should not analyze more than 70 slides and not work more than 7 hours per day. Cytopathologist must not have fatigue and should concentrate to prevent diagnostic error (Bengtsson et al., 2014). The task is also challenging for cytopathologists due to the presence of cervical cells with spurious edges, overlapping cells, neutrophils, and artifacts in the slide and in the field of view after staining. Hence, a computational approach is necessary towards its early detection (Jantzen et.al., 2005). The purpose of implementing digital cell analysis for Pap Smear evaluation is to prevent human errors, increase efficiency, significantly reduce test time, and detect cancer in its early stage. Previous studies using a generated dataset show that nuclear shape features from images have discrimination power to computationally identify normal and abnormal cells (Bhowmik et.al., 2018). In this study, thirteen nuclear features were quantified and used to differentiate the cell

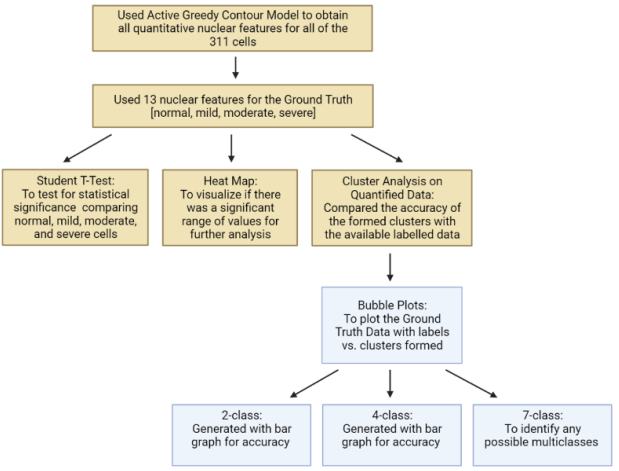
images into two class and multiclass data based on their nuclear features from a benchmark dataset. The data was then compared for a visual analysis of the cells by graphing varying relationships between the features using a software called Tableau. One of the main analysis methods used is called clustering and is based on the K-means algorithm which can group the data into one class based on its relativity to the average of that cluster. The objective is to quantify thirteen nucleus shape features to differentiate normal and abnormal cells, and to examine potential multiclasses among the abnormal cells.

Materials & Methods

Contouring the Nucleus and Shape Features Examined

The cervical cancer cell images, the 389 cells that we began with, were taken from the Herlev University Pap Smear dataset (Jantzen et al., 2005) and served as the Ground Truth for analysis of nuclear shape characteristics. The cell images were created utilizing a digital camera microscope and cytotechnicians classified each cell into three classes of normal (superficial squamous epithelia, intermediate squamous epithelia, columnar epithelial) and four classes of abnormal (mild squamous non-keratinizing dysplasia, moderate squamous non-keratinizing dysplasia, severe squamous non-keratinizing dysplasia, squamous cell carcinoma) on the basis of the morphological features (Bhatt, et al, 2021). In this study, 92 normal cervical cancer cell images were selected randomly from the three categories and 297 abnormal cervical cancer cell images were selected comprising of 105 abnormal-mild squamous non-keratinizing dysplasia cells, 92 abnormal-moderate squamous non-keratinizing dysplasia cells and 100 abnormal-severe squamous non-keratinizing dysplasia cells. A flow diagram of the analysis conducted in this study is shown in Fig. 1. The Greedy Active Contour Model was used for segmentation of

the nucleus of the cells (Williams and Shah, 1992). The nucleus is the region of interest and by contouring the circumference of the nucleus, 13 distinctive features were quantified by using MATLAB. The 13 features are: Nuclear Area, Nuclear Perimeter, Solidity, Nuclear Roundness, Major Axis Length, Minor Axis Length, Extent, Equivalent Diameter, Eccentricity, Elongation, Convex Area, maximum length from center of gravity to perimeter (MGP) and average length from center of gravity to perimeter (AGP). Each feature is defined by a mathematical model. The nuclear region of the pre-cancerous cells are larger and sometimes misshaped when compared to normal cells and by finding the differences in shape, shape-based classification can be done. Sample images and nucleus contouring are shown in Fig. 2 (normal cell) and Fig. 3A and 3B (two abnormal cells). The quantified data was then input into an excel file which was generated to compare with the ground truth data obtained from the Herlev dataset. Normal versus any category of abnormal feature was considered statistically significant if by Student's T-Test there was $p < 0.05$.



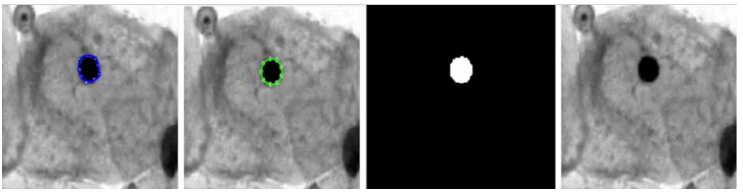


Figure 1: Flow diagram of the analysis conducted.

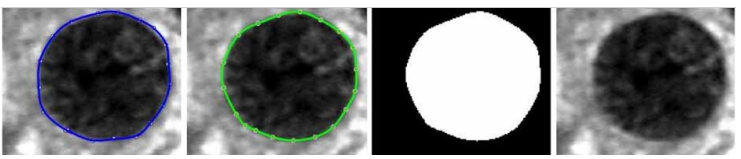


Figure 2: Normal Cell Initialization of Contour Points, Active Contour Model, Segmented Image

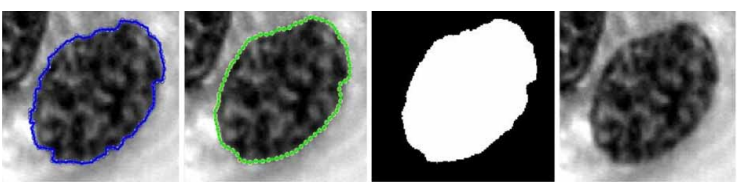


Figure 3A: Abnormal Cell Initialization of Contour Points, Active Contour Model, Segmented Image

Figure 3B: Another Sample Abnormal Cell Initialization of Contour Points, Active Contour Model, Segmented Image

Tableau and Data Visualization

The visualizations were generated using Tableau, a visual analytics software that allows for datasets to be compared through one or more distinctive features. Here 311 cells were analyzed using visualization techniques such as bubble graphs and heat maps, and classified into normal, mild-abnormal, moderate-abnormal, and severe-abnormal

labels. Features were also visualized against other features in a one versus one graphical analysis, as well as multiple features against each other. Labeled graphs served as the positive controls within this research also known as the Ground Truth as shown in Fig. 5A and 6A. The Ground Truth refers to the known information about the data source, its labeled definitions, and marked label annotations.

Clustering and K-means Algorithm

Once the Ground Truth was established, an unsupervised machine learning technique was utilized, called clustering. Unsupervised learning is a machine learning technique in which the software does not supervise the model or use any predetermined labels, rather it discovers patterns and organizes the data into various clusters using Tableau's inbuilt clustering analysis feature. Where k represents the K cluster centers, u_k represents the k_{th} center, and x_i represents the i_{th} point in the dataset. This equation takes all of the variables in relation to the distortion of the clusters which is the sum of the squared distances of points from the cluster centers:

The K-means clustering algorithm is a technique that splits the

$$d = \sum_{k=1}^{k} \sum_{i=1}^{n} ||x_i - u_k||^2$$

data into K number of clusters where every cluster has a center similarity that is the mean value of all the points within that given cluster. It is an iterative scheme that evolves K crisp and compact clusters in the data such that a measure is minimized. Essentially, the K-means algorithm groups the data together based on a similarity between the data points.

Results

Quantification of Nuclear Features for 311 Cells and Heat Map

The dataset had 389 cells. However, there was several overlapping data regarding nuclear features of moderate or severe-abnormal, and therefore, for this particular study, those cells were removed and 311 cells were used [89- normal; 222-abnormal]. The thirteen nuclear shape features quantified using MATLAB were Nuclear Area, Nuclear Perimeter, Solidity, Nuclear Roundness, Major Axis Length, Minor Axis Length, Extent, Equivalent Diameter, Eccentricity, Elongation, Convex Area, MGP and AGP. A statistical analysis was conducted to obtain average and standard deviation for each feature and each category, as shown in Table 1. Upon analysis of these quantified shape features in the dataset, we found nine features to be discriminatory by the Student T-Test, because their p-values were less than 0.05 (normal vs any class of abnormal). Now with the quantified features, we proceeded to visualize all the data using Tableau by examining various combinations of five features with discriminatory value (p-value < 0.05, statistically significant) and two non-discriminatory (p-value > 0.05, not significant) features.

	Normal	Mild	Moderate	Severe
Nuclear Area	233.75 ± 128.05	2517.08 ± 637.08 **(8.51E-69)**	4230.54 ± 1186.79 **(1.91E-72)**	8969.38 ± 1822.37 **(1.03E-94)**
Extent	0.72 ± 0.05	0.73 ± 0.5 (0.7636)	0.71 ± 0.07 (0.0616)	0.73 ± 0.05 (0.7527)
Nuclear Roundness	0.49 ± 0.04	0.53 ± 0.03 **(6.22E-09)**	0.52 ± 0.07 **(0.0068)**	0.55 ± 0.04 **(2.55E-13)**
Nuclear Perimeter	74.22 ± 19.23	241.93 ± 33.78 **(3.25E-78)**	320.24 ± 48.82 **(7.26E-95)**	453.05 ± 45.28 **(1.05E-127)**
Solidity	0.95 ± 0.02	0.97 ± 0.01 **(8.33E-08)**	0.97 ± 0.03 **(0.0020)**	0.98 ± 0.01 **(5.37E-18)**
Equivalent Diameter	16.60 ± 4.55	56.14 ± 7.39 **(5.51E-81)**	72.67 ± 10.30 **(7.26E-99)**	106.35 ± 10.59 **(1.79E-128)**
Eccentricity	0.63 ± 0.16	0.65 ± 0.18 (0.6019)	0.68 ± 0.14 **(0.0243)**	0.64 ± 0.14 (0.6494)
MGP	62.69 ± 11.14	69.93 ± 12.91 **(0.0004)**	76.29 ± 15.51 **(4.18E-10)**	81.57 ± 12.83 **(2.14E-19)**
AGP	62.32 ± 11.23	65.84 ± 13.56 (0.0883)	70.47 ± 16.94 **(0.0002)**	70.35 ± 14.45 **(7.70E-5)**
Elongation	0.75 ± 0.14	0.73 ± 0.13 (0.5303)	0.70 ± 0.13 **(0.0461)**	0.74 ± 0.12 (0.9545)
Major Axis Length	19.72 ± 5.91	66.85 ± 11.03 **(1.04E-70)**	88.12 ± 13.13 **(1.09E-95)**	125.03 ± 14.37 **(8.51E-119)**
Minor Axis Length	14.35 ± 4.11	48.01 ± 7.53 **(5.91E-73)**	61.23 ± 11.30 **(1.52E-82)**	92.04 ±12.89 **(1.54E-107)**
Convex Area	243.07 ± 132.11	2594.75 ± 660.40 **(5.06E-69)**	4365.83 ± 1212.05 **(1.18E-73)**	9158.68 ±1853.2 **(1.09E-95)**

Table 1: Mean ± Standard Deviation of all thirteen nuclear features extracted from images of 311 cells. P-values for normal vs. each abnormal class are shown in the parentheses and statistically significant ($p < 0.05$) data is in bold. Nuclear perimeter, equivalent diameter, major axis length are measured in micrometer while others are ratios.

Creation of Heat Map Using Seven Features

We generated the heat map with seven features as shown in Fig. 4. The seven features are shown in columns, 311 cells are in the rows, and the color gradient represents the range of values for the nuclear features. We found that the Nuclear Area, Equivalent Diameter, Major Axis Length and Minor Axis Length showed a clear distinction between the normal, and abnormal-mild, abnormal-moderate and abnormal-severe cells whereas Extent and Elongation showed no clear gradient because the range of values were not significant to distinguish the classes of cells. MGP feature was statistically significant when comparing all three abnormal cell types versus normal (Table 1) and the heat map shows a marginal gradient for MGP among the 311 cells.

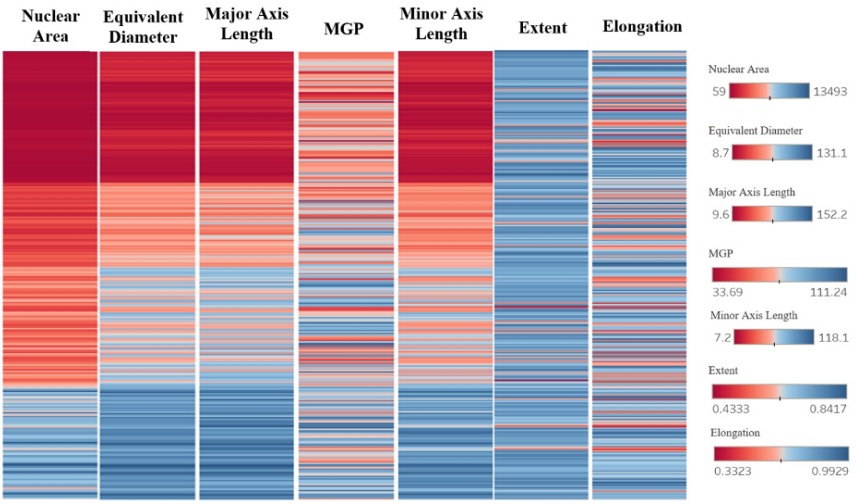


Figure 4: Heat Map for normal, mild-abnormal, moderate-abnormal, and severe-abnormal cells for seven features.

Bubble Plots Provided the Visualization of Two Clusters

A two-class bubble graph was obtained using the quantified four features (MGP versus Equivalent Diameter, Major Axis Length and Extent) that were input and were pre-labeled into two categories: normal and abnormal (Fig. 5A). Next, the dataset was input without providing labeled categories such as normal or any class of abnormal and organized it into 2 clusters using the k-means algorithm which is the Ground Truth data (Fig. 5B).

When we compared the labeled bubble graph with the clustered bubble graph (compare Fig. 5A vs 5B), we identified an overlap between the normal and abnormal cells. The success rate which is defined by the true and false rate represents all of the cells that matched the Ground Truth data, or did not, respectively, was found to be 89% true- and 10.51% as false rate (Fig. 5C).

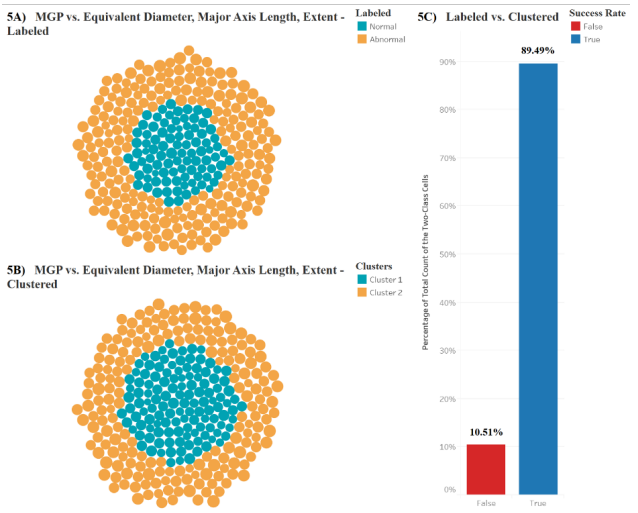


Figure 5: Two Class Bubble Graph. A is Ground truth that was labeled; B is unlabeled and categorized using K-Means Algorithm; C is success rate.

Bubble Plots Show the Possibility of 4 or More Clusters

A four-class bubble graph was generated using the four different features above (Equivalent Diameter vs. Extent, MGP, Nuclear Area). Figure 6A shows the labeled normal, mild, moderate and severe abnormal cells bubble representation, considering all the 4 features as ground truth data. We obtained 4 categories as expected. Next, the dataset was input without providing labeled categories of normal or any class of abnormal (Fig. 6B) and by the K-means algorithm we obtained 4 clusters, but when comparing Fig. 6B to 6A, the success rate was only 78% as true rate and approximately 22% as false rate.

By examining 2 clusters and 4 clusters, the accuracy or success rate was reduced from 89% to 78%. This means that the quantified values may be so close that there is overlap for some features and it is not possible to distinguish between them.

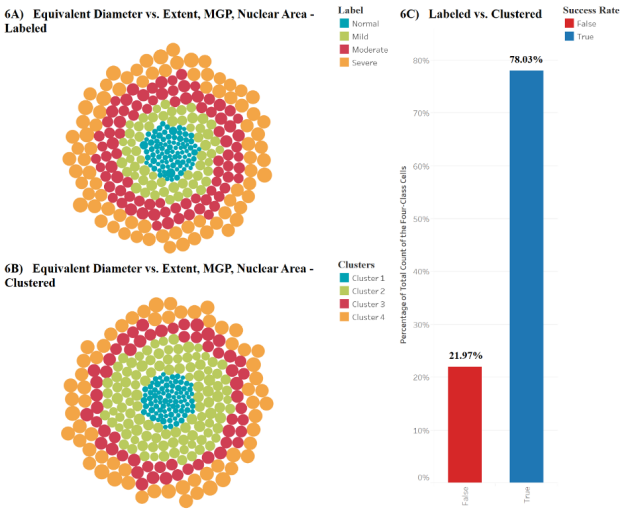


Figure 6: Four Class Bubble Graph. A is Ground truth that was pre-labeled; B is unlabeled and categorized Using K-Means Algorithm; C is success rate

We examined the possibility of the existence of multi classes i.e. more than 4. Different features and/or combinations of features were input. By comparing nuclear area to equivalent diameter, elongation and minor axis length, and without any labels of normal or abnormal classes, we obtained 7 clusters (Fig. 7) that were much more organized by size of bubble as compared to 4-class bubble plot of Fig. 6B. Therefore, using the K-means algorithm it can be concluded that there are multiple classes identified within the normal and abnormal category. The numbers in each bubble represent the nuclear areas of each cell.

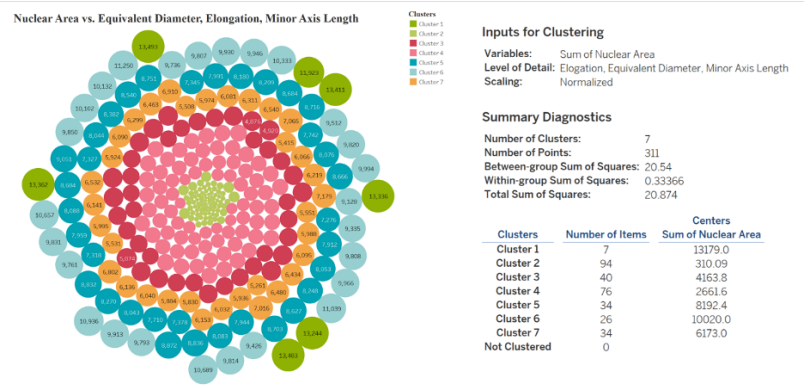


Figure 7: Seven Classes Bubble Graph and Summary of the Clusters Formed

Discussion

Our findings suggest that 7 out of the 13 measured parameters showed a significant difference between an abnormal and a normal cell nuclei, through statistical analysis and visual representation. Combining discriminatory features such as Nuclear Area, Equivalent Diameter, Major Axis Length, MGP and Minor Axis Length, along with non-discriminatory features such as Elongation and Extent,

plays an important role in the formation of multiple clusters and gives a good success rate value.

Using the K-means clustering analysis in Tableau, we were able to successfully distinguish the dataset between a normal and abnormal cell and showed an 89.49% success when we compared the two. The K-means algorithm also identified multiple classes within the non-labeled dataset within the abnormal class. This means there is the possibility that 'intermediate' classes such as extreme-severe abnormal cells and mild-moderate abnormal cells exist in this dataset. Indeed, there exist many categories of normal (Superficial squamous epithelial, Intermediate squamous epithelial, Columnar epithelial) and also of abnormal (Mild squamous non-keratinizing dysplasia, Moderate squamous non-keratinizing dysplasia, Severe squamous non-keratinizing dysplasia, Squamous cell carcinoma in situ intermediate) (Norup, 2005; Mbaga and Zhijin, 2015). For comparison with other studies, we found two-class classifier results with the Herlev Dataset of PAP smear images that characterized nucleus and cytoplasm features using support vector machine (Bora et al., 2017; Sajeena and Jereesh, 2015) and they found accuracies of 95% and 93%, respectively. However, we did not find any reports of clustering or multiple classes identification as we have performed.

Future studies include identification of the 7 clusters found in the bubble graph in Fig. 6 and confirm the presence of 1 cluster of normal cells and the rest of abnormal cells. Another extension of this research is to collect data from other larger datasets and apply the K-means algorithm to increase the consistency and accuracy of identifying the cell type. K-means clustering algorithm can be implemented in an application or software which can be utilized in laboratories for digital pathology analysis where the cells obtained from a PAP smear are compared to normal and abnormal cells as defined based on the Ground Truth nuclear characteristics implemented in the software. Upon collection of cells, they can be

imaged and the program can measure its nuclear characteristics and diagnose whether the cell is abnormal and to what degree of abnormality it may exhibit.

A limitation of this study is that we have only quantified 311 cells for their nuclear features, however we want to work towards quantifying features for more cells such as 1000 cells and to minimize or diminish the feature overlap obtained through the clustering method and maximize accuracy or the success rate.

Conclusion

By using a benchmark dataset, we quantified thirteen nucleus shape features to differentiate normal and abnormal cells and found potentially seven classes among the normal and abnormal precancerous cervical cells by visualization techniques. This deserves further study for the advancement of the methods of detection of cervical cancer.

Acknowledgement: We thank Ms. SaiSrija Edara, MS student in Data Science, for additional mentorship and directions in this project, and for providing feedback on this manuscript.

References

American Cancer Society. Cancer Statistics. (Accessed Nov 1, 2021). www.cancer.org.

Arbyn, M., Weiderpass, E., Bruni, L., de Sanjosé, S., Saraiya, M., Ferlay, J., & Bray, F. (2020). Estimates of incidence and mortality of cervical cancer in 2018: a worldwide analysis. The Lancet. Global health, 8(2), e191–e203. https://doi.org/10.1016/S2214-109X(19)30482-6

Bengtsson, E., & Malm, P. (2014). Screening for cervical cancer using automated analysis of PAP-smears. Computational and mathematical methods in medicine, 2014, 842037. https://doi.org/10.1155/2014/842037.

Bhatt, A.R., Ganata, A., & Kotecha, K. (2021). Cervical cancer detection in pap smear whole slide images using convNet with transfer learning and progressive resizing. *PeerJ Computer Science*, 7:e348.

Bhowmik, M.K., Roy, S.D., Dutta, A., & Nath, N. (2018). Nucleus region segmentation towards cervical cancer screening using AGMC-TU pap-smear dataset. Proceedings of the International Conference on Pattern Recognition and Artificial Intelligence. Union, N.J., 44-53.

Bora, K., Chowdhury, M., Mahanta, L. B., Kundu, M. K., & Das, A. K. (2017). Automated classification of Pap smear images to detect cervical dysplasia. Computer methods and programs in biomedicine, 13*8*, 31–47. https://doi.org/10.1016/j.cmpb.2016.10.001

Jantzen, J., Norup, J., Dounias, G., & Bjerregaard, B. (2005). Pap-smear Benchmark Data for Pattern Classification. In Proc. NiSIS 2005: Nature Inspired Smart Information Systems, pp. 1-9.

Mbaga, A., & Zhijun, P. (2015). Pap Smear Images Classification for Early Detection of Cervical Cancer. International Journal of Computer Applications, Vol. 118, 10-18.

Norup J. (2005). Classification of Pap-smear data by transductive neuro-fuzzy methods. Master's thesis, Tech Univ Denmark Oersted-DTU;71.

Sajeena, T. A. & Jereesh, A. S. (2015). Automated cervical cancer detection through RGVF segmentation and SVM classification. 2015 International Conference on Computing and Network Communications (CoCoNet), pp. 663-669, doi: 10.1109/CoCoNet.2015.7411260.

Williams, D.J., & Shah, M. (1992). A fast algorithm for active contours and curvature estimation. CVGIP: Image understanding, Vol.55, no.1, pp. 14-26.