Image Analysis of Human Cervical Cells for Nucleus Features: Towards Early Detection of Cervical Cancer

NEW YORK INSTITUTE OF TECHNOLOGY

Tanvi Patel, Caroline Fernandez, Brian Siuni, Krushang Pandya, Sai Srija Edara[1] and Dr. Niharika Nath

Department of Biological & Chemical Sciences; [1]Department of Computer Science, New York Institute of Technology, New York, NY

Introduction

Cancer is the second leading cause of death worldwide, accounting for 14% of all deaths, in the Americas, Europe, and the Western Pacific regions. Cervical cancer is the fourth most commonly occurring cancer in women and the eighth most commonly occurring cancer overall. There were over 500,000 new cases in 2018. Papanicolaou Smear Test is a commonly used method to detect the presence of precancerous and cancerous cells. 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. Moreover, most of the Cervical Cancer cases are reported in developing and underdeveloped countries where the resources are limited. Hence, a computational approach is necessary towards the early detection of Cervical Cancer. In this study, thirteen (13) nuclear features were quantified and used to differentiate the cell images into two class/multiclass data based on their nuclear features. 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.

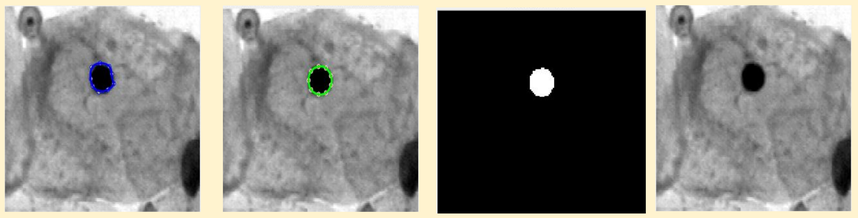
Objective

The objective of this project is to quantify thirteen nucleus shape features to determine how the various shape features could differentiate normal and abnormal cells, and to examine the accuracy of prediction of classification of these cells into normal and abnormal.

Materials & Methods

The cervical cancer images taken from the Herlev University Pap Smear dataset served as the Ground Truth for analysis of nuclear characteristics. The greedy Active Contour Model (ACM) was used to isolate the nucleus of the cells. The nucleus tends to be the region of interest and by contouring the circumference of the nucleus, 13 distinctive features were quantified by using MATLAB. The nuclear region of the pre-cancerous cells are larger and misshaped when compared to normal cells and by finding the differences in shape, shape-based classification can be done. The quantified data was then input into an excel file which was generated to compare with the ground truth data obtained from the norup dataset. The visualizations were generated using Tableau by comparing one or more distinctive features and different patterns were observed which supported the hypothesis of differentiating between normal and subclasses of abnormal.

Figure 1: Normal Cell



` Initialization of Contour Points, Active Contour Model, Segmented Image

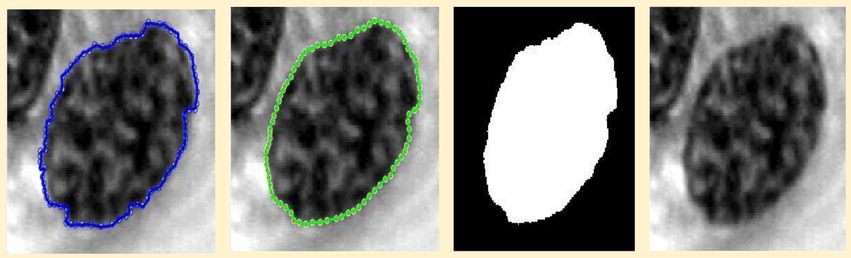


Figure 2: Abnormal Cell

13 Nuclear Shape Parameters

- Convex Area
- Eccentricity
- Elongation
- Extent
- MGP
- Nuclear Area
- Nuclear Perimeter
- Nuclear Roundness
- Solidity
- AGP
- Major Axis Length
- Minor Axis Length
- Equivalent Diameter

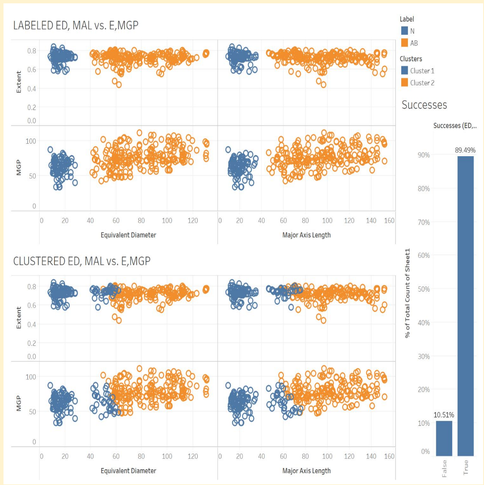
Results

Table 1: Mean ± Standard Deviation of six distinctive features identified. Patterns can be determined by analyzing the variation in values amongst normal, mild, moderate and severe cells.

	Normal	Mild	Moderate	Severe
Nuclear Area	238.79 ± 130.61	2517.08 ± 637.08	4230.54 ± 1186.79	8969.38 ± 1822.37
Nuclear Perimeter	75.09 ± 19.57	241.93 ± 33.78	320.24 ± 48.82	453.05 ± 45.28
Equivalent Diameter	16.82 ± 4.63	56.14 ± 7.39	72.67± 10.30	106.35 ± 10.59
Major Axis Length	19.96 ± 6.04	66.85 ± 11.03	88.12 ± 13.13	125.03 ± 14.37
MGP	63.02 ± 11.11	69.93 ± 12.91	76.29 ± 15.51	81.57 ± 12.83
Extent	0.72 ± 0.05	0.73 ± 0.5	0.71 ± 0.07	0.73 ± 0.05

Figure 3: Heat Map
The heat map consisting of six features in the columns, 315 cells in the rows and the color gradient represents the range of values for the nuclear feature. Nuclear Area, Nuclear Perimeter, Equivalent Diameter and Major Axis Length showed a clear distinction between normal, mild, moderate and severe cells whereas The range of values for MGP and Extent were not significant to distinguish the classes of cells.

Figure 4: Scatter Plot
When comparing the clustered scatter plots to the known Ground Truth data, the clustered scatter plots showed an 89.49% accuracy in differentiating between the Normal and Abnormal cells. Overall, by using Tableau's clustering analysis through the k-means algorithm, the non-labeled dataset was successfully clustered into normal (N) and abnormal (AB) cells.



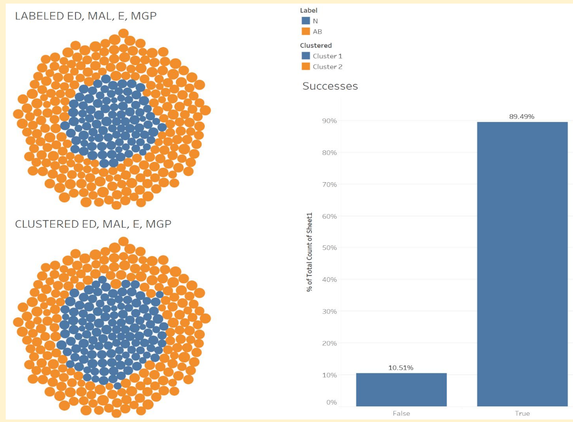


Figure 5: Bubble Graph - Two Class
When comparing the clustered labeled packed bubble graphs to the known Ground Truth data, the clustered bubble graphs also show an 89.49% accuracy when differentiating between the Normal and Abnormal cells. By using a different method of visualization: packed bubble graph, we can visualize how the normal (N) and abnormal (AB) cells are differentiated without labels using the K-means algorithm.

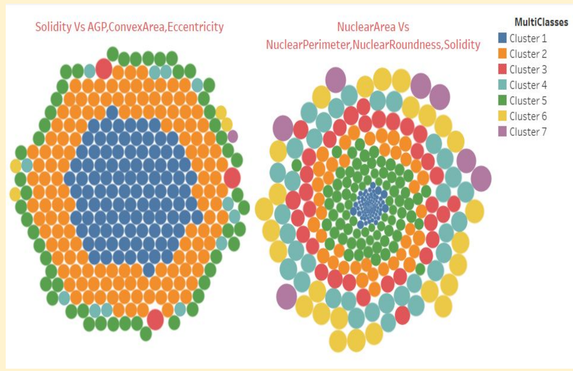


Figure 6: Bubble Graph - MultiClass
When comparing the clustered labeled packed bubble graphs to the known Ground Truth data, multiple classes are identified. Using the K-means clustering method in Tableau, it can be concluded that there are multiple classes identified within the abnormal (AB) category through the 7 different clusters formed with several features.

Discussion

- Out of the 13 measured parameters, 6 parameters showed significant differences between abnormal and normal cell nuclei.
- Nuclear Area, Nuclear Perimeter, Equivalent Diameter, Major Axis Length, MGP and Extent were the six features found to be significant through the visual analysis using Tableau.
- Through the K-means clustering analysis in Tableau, the dataset was successfully distinguished between normal and abnormal cells and gave an 89.49% success percentage when compared between the two. In addition, the K-means clustering algorithm was also able to identify multiple classes within the non-labeled dataset within the abnormal class.
- Future studies may include identification of 7 clusters found in the bubble graph in Figure 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 different databases and apply the K-means clustering algorithm of Tableau to develop a consistency in result and to increase the accuracy of identifying the cell type.

References

(1)Bhowmik, M. K., Nath, N., Dr., Datta, A., & Ghosh, A. K. (2017, December). Shape Feature Based Automatic Abnormality Detection of Cervico-Vaginal Pap Smears. Retrieved June 6, 2019, from http://www.joig.org/uploadfile/2018/0323/20180323015725353.pdf

(2)Jantzen J, Norup J, Dounias G, Bjerregaard B. Pap-smear benchmark data for pattern classification. In: Proc NiSIS 2005 Nat inspired Smart Inf Syst. p. 1–9, 2005.

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

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

(5)Uyar, D. & Rader, J. Genomics of cervical cancer and the role of human papillomavirus pathobiology. *Clin. Chem.* **60**, 144–146 (2014)