A Neural Network Detects Abnormal Pap Smear Images

Wei Hung Hsu
Harvard T.H. Chan School of Public Health
677 Huntington Ave, Boston, MA 02115
morris754@gmail.com

Maja Garbulinska
T.H. Chan School of Public Health
677 Huntington Ave, Boston, MA 02115
majagarbulinska@gmail.com

Abstract

Cervical cancer is a very common cancer in women worldwide, but it can be easily prevented by either the HPV vaccine or frequent pap smear screening. Most of the deaths from cervical cancer occur in low income countries where screening is less available. Even in more developed countries, trained specialist have to face a taxing task of looking at many slides under a microscope just to find a few that are abnormal. Deep learning presents great opportunities for revolutionizing abnormality detection on pap smear slides. In this project we present the result of four experiments we performed with neural network models trained, validated and tested with the SIPaKMeD pap smear data set. Our models classified respective images with high accuracy levels and could be used to design real life applications that would highly reduce the costs of pap smear screenings and possibly reduce cervical cancer rates.

1. Introduction

1.1. Cervical Cancer

Cervical cancer is the third most common cancer worldwide in women leading to thousands of deaths every year. Even though, if untreated, cervical cancer is fatal, it is very easy to prevent. Infection with a high-risk human papillomavirus (HPV) strain is necessary but not sufficient for the cervical cancer to develop. One of the way of preventing this type of cancer is therefore the HPV vaccine. This vaccine however, is relatively new and many women might not have had the chance to get it before they became infected. Another way of preventing cervical cancer is frequent screening with a pap smear test. The pap smear tests are an easy way of detecting lesions and abnormal cells on the cervix. Most of the deaths from cervical cancer occur in developing countries where screening is not as well available as in most of the developed economies. [1]

1.2. Pap Smear Evaluation

During a medical visit, a sample of cells from the cervix is scraped by a medical professional using a special brush and sent to a lab for evaluation. The abnormal results can be either benign lesions that should be monitored, more serious changes that if untreated could lead to cancer, or already more serious changes that are dangerous and require intermediate medical intervention. [6]

Currently, the pap smears specimens are evaluated by technicians or pathologists who look at the cells under a microscope and report their findings. The procedure is very tedious and demanding leading to high costs. The specialist looking at pap smear results have to be able to identify even a small number of abnormal cells among a few hundred thousands normal cells. Due to the taxing nature of this task, a specialist can only look at around 100 specimens per day. [5, 4]

1.3. Motivation for an Automated Detection System

Introducing a good and cost-effective automated detection system for abnormal pap smears could result in lower mortality as screening would be cheaper and available. Around 7% to 10% of all pap smears are assessed to be abnormal. It takes 5-10 minutes on average for a specialist to evaluate a pap smear under a microscope. Assuming that an algorithm could accurately detect 20% of the pap smears that look suspicious, the other 80% would not even have to be evaluated by a human. This solution would lead to incredible cost savings.

An estimated 85% of cervical cancer cases occur in low-income countries. These countries face shortages of specialists and need cheap screening methods to reduce mortality. Introducing deep learning models in this field could be a great first step towards this goal. [2, 6, 4, 12]

1.4. Deep Learning for Images

Deep Learning and especially convolutional neural networks (CNNs) for image classification have been growing in popularity. Efficient processing using better Graphics

Processing Units allows neural networks to learn key features in images fast by adjusting model parameters step by step to minimize a specified loss function. CNN have proven to be efficient and good at image classification and object detection. [3] [**?**]

1.5. Previous Work

Various projects have been proposed to commercialize automated detection of abnormal pap smears over the past decades but the cost-savings that have been achieved are not sufficient. System like "BD Focal Point GS Imaging System" have been approved by the Federal Drug Administration in the United States, but only a maximum of 25% of slides are classified as "No further review". The system also provides the possibility of visually reviewing fields of interest.

Despite the fact that these systems are available in countries like the United States, they do not have a significant impact in low-income countries due to the high costs involved in their purchase and maintenance.

Currently, there are no commercialized applications that use a deep learning algorithm to classify pap smear images and little work has been done on the topic. Some researchers achieve good results while classifying single cells instead of the whole image. This approach however, is not very cost effective as the task of cropping the cells out of the image can itself be very time consuming. Great results have been also recently achieved with Waka classifiers on full images. [9, 12, 10]

2. Approach

2.1. Data

This project uses a publicly available image dataset SIPaKMeD, which consists of 4049 annotated cells images. The cells were classified by expert cytopathologists into five different classes/categories. The categories are:

- superficial-intermediate (normal)
- parabasal (normal)
- dyskeratotic (abnormal)
- koilocytotic (abnormal)
- metaplastic (benign)

The cells were cropped from 966 full pap smear cluster cell images and these images are also provided with the data set. Table 1 presents how many images of cell clusters and cropped cells were provided for each of the five categories. And Figure 1 shows examples of cropped cell images for each of the five categories and Figure 2 shows examples of full images.

For part of our analysis, we classified the images further into just two categories. "abnormal" which refers to images that should be seen by a specialist, and "normal". The idea of this project, given that the vast majority of pap smears are do not have abnormal cells, is to find as many "normal" images as possible, so that specialist do not have to look at them while still not missing abnormal cases. "Superficial-intermediate", "metaplastic" and "parabasal" categories were merged into "normal" and the rest into "abnormal"

Table 1. Number of images per category.

Category	Full Images	Cropped Cells
superficial-intermediate	126	813
parabasal	108	787
koilocytotic	238	825
metaplastic	271	793
dyskeratotic	223	813

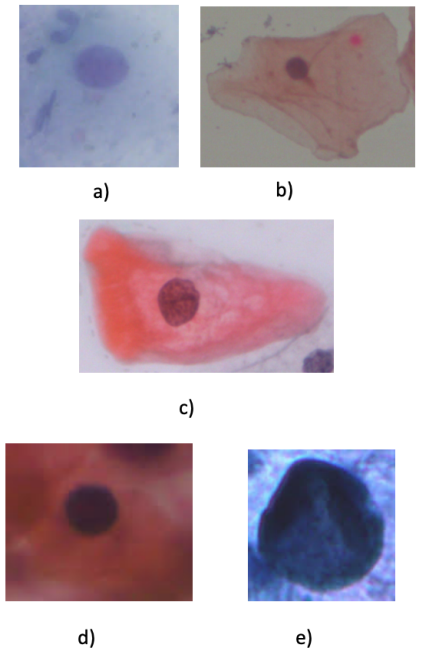

Figure 1. Examples of cropped cells: a) parabasal, b) superficial-intermediate, c) koilocytotic, d) dyskeratotic, e) metaplastic

2.2. Methods

2.2.1 Four Different Experiments

We started this project by using the full images to train a network and predict the five different classes described above. We then trained the model on cropped cell images.

Because of suitability for real life applications we wanted to go further and have an algorithm that would classify full images as normal or abnormal. We also expected to know if more images per class would improve accuracy.

Another experiment we did was to train the network on

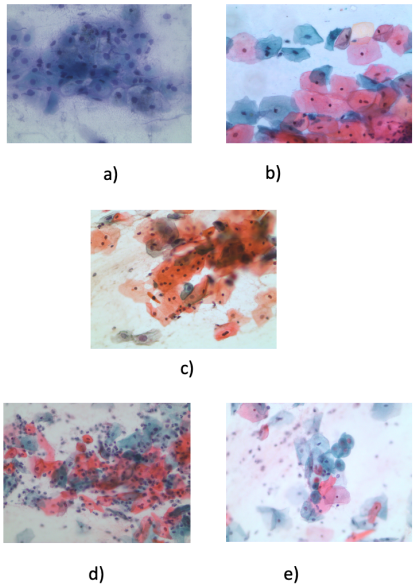

Figure 2. Examples of full images: a) parabasal, b) superficial-intermediate, c) koilocytotic, d) dyskeratotic, e) metaplastic

cropped cells but categorized as only to classes ”normal” and ”abnormal”.

As mentioned before in our experiment, metaplastic cells are assumed to be ”normal”. Hence normal cells include: superficial-intermediate, parabasal and metaplastic. Abnormal cells include dyskeratotic and koilocytotic.

2.2.2 The Models

To train the models we divided the data into three different sets: training, validation, test with the following proportions 0.8/0.1/0.1.

We used several convolutional neural network (CNN) models. We tried VGG-11, Resnet-18, Resnet-50 and Resnet-101. The input to the networks are raw RGB images. They are resized to 224 x 224 pixels. To increase accuracy, we applied data augmentation by randomly rotating the training images. Next we added Pytroch's AdaptiveMaxPool and AdaptiveAveragePool2d, flattened them out and concatenated them to form a linear layer. Then we applied a dropout layer with a dropout rate of 0.5 followed with a linear layer.

The optimizer Adam was used to train the models (batch size 16, learning rate 10^{-3}). To make the models converge faster and increase accuracy, we implemented a scheduler which reduces the learning rate at each epoch. Training is terminated after 50 epochs.

All the models were trained using Pytorch for Python. Our code is ready to be shared. [7, 8, 11].

3. Results

3.1. Achieved Accuracy

We have trained the models on 80% of the images, using additional 10% as validation. We then tested the models on the remaining 10% of the images. The presented classification. The classification accuracies are presented in Table 2. As you can see in the first column, we include the number of classes the data set was divided(two vs five) into as well as the type of images(cropped vs full). For all of these four approaches we trained a VGG, a Resnet18, a Resnet50 and, a Resnet101.

Table 2. Classification Accuracy for Different Methods

	VGG	Resnet18	Resnet50	Resnet101
5 class/ full image	0.470	0.823	0.784	0.803
5 class/ cropped image	0.887	0.933	0.943	0.899
2 class/ full image	0.734	0.785	0.826	0.806
2 class/ cropped image	0.926	0.945	0.933	0.933

3.2. Five Classes Approach Results

Table 3.2 shows the results of classification full images using Resnet50. The koilocytotic cells are the most challenging cells to be distinguished correctly in a full image. The koilocytotic slides are often classified as metaplastic or dyskeratotic.

Resnet50 classification results on full images

Predict \ Truth	Dysk.	Koi.	Meta.	Para.	Super.
Dyskeratotic	0.869	0.13	0.0	0.0	0.0
Koilocytotic	0.0	0.68	0.12	0.12	0.08
Metaplastic	0.035	0.107	0.714	0.035	0.107
Parabasal	0.0	0.0	0.0	0.916	0.083
Superficial	0.0	0.071	0.0	0.071	0.857

Table 3.2 shows the results of classification cropped cell image using Resnet50. The metaplastic cells have the lowest accuracy. Parabasal have the highest probability to be classified correctly. In fact, in our test set, all of the parabasel cells were classified correctly.

Resnet50 classification on cropped images

Predict \ Truth	Dysk.	Koi.	Meta.	Para.	Super.
Dyskeratotic	0.951	0.036	0.012	0.0	0.0
Koilocytotic	0.012	0.939	0.036	0.0	0.012
Metaplastic	0.012	0.037	0.875	0.05	0.025
Parabasal	0.0	0.0	0.0	1.0	0.0
Superficial	0.0	0.035	0.035	0.0	0.928

3.3. Two Classes Approach Results

Table 3.3 and table 3.3 present confusion matrices for Resnet50 classification test results for full images and cropped images respectively.

The results look good given the fact that the data set is not very big.

It is important to note that these results relay on the default 50% threshold, meaning that an image is classified as abnormal when the probability of it being abnormal estimated by the model is greater than 50%. This threshold however might not be appropriate in the real world clinical setting where false positive results should be avoided.

Resnet50 classification on full images

Predict \ Truth	Normal	Abnormal
Normal	0.803	0.196
Abnormal	0.148	0.851

Resnet50 classification on cropped images

Predict \ Truth	Normal	Abnormal
Normal	0.962	0.037
Abnormal	0.103	0.896

4. Additional features

For models with two classes the class is assigned to an image based on the probability of this image being of a given class. The default threshold is 50%. For some applications, such as the problems faced in themedical field however where false negative results should be avoided, the threshold can be readjusted.

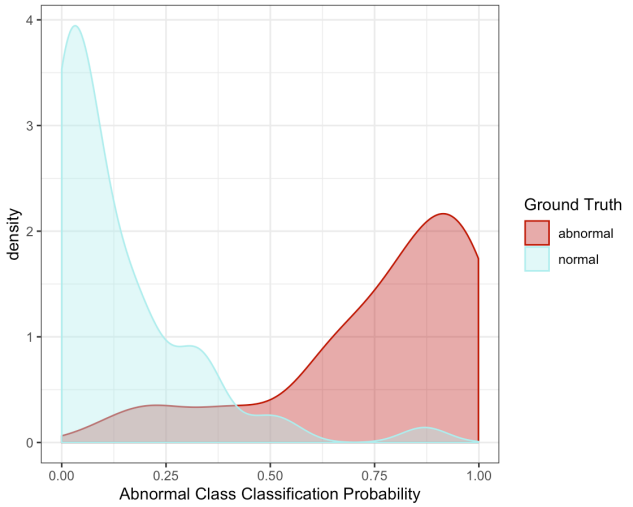

Figure 3. Density of the probabilities of an image being classified as abnormal depending on the real class.

Figure 3 shows the distribution of the "abnormal" class probabilities for full images that are either normal or abnormal. The ground truth abnormal images are displayed in red and the ground truth normal images are displayed in blue.

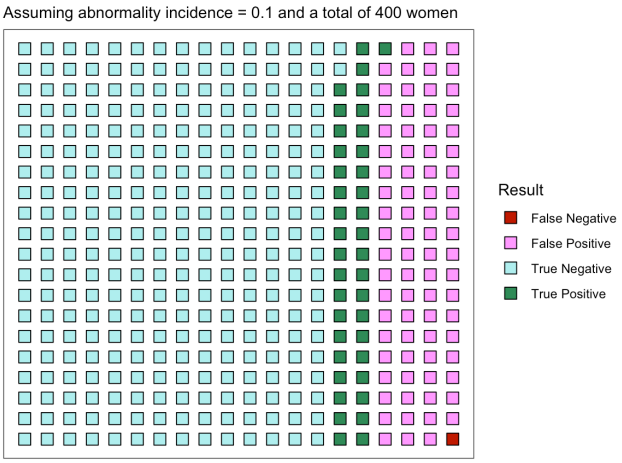

Figure 4. Predicted number of pap smears with each result.

We can clearly see that there is a huge difference in the distributions, but there are still images that are normal but have a high probability of being classified as abnormal and vice-versa.

In addition to our prediction model, we wrote an algorithm that takes the predicted probabilities for each class and adjusts the classification probability threshold.

Another extension we wrote is a function that calculates false positive rate, false negative rate, true positive rate and true negative rate based on the probabilities resulting from the model and a specified threshold.

Figure 4 has been automatically generated using our function and shows the estimated number of pap smears results that are false negative, false positive, true negative, and true positive. We used output from the Resnet50 model trained on full images.

We assume a potential set of 400 images displayed in this graph as squares. We applied the probability threshold to be 20% meaning, if an image was predicted by the network to have above 20% probability of being abnormal, we classify it as abnormal. We assumed that 10 % of the pap smears in this imaginary lab are abnormal.

The true positive rate with this setting is 98%, which means that of all abnormal images, the algorithm detects 98%.

The false positive rate was 21%, meaning that of all normal images we misclassify 21%. In the simulated cohort of 400 women, 78 would be told that their pap smear results were abnormal. Even though, a false positive result can lead to stress, these women would probably asked to come and see a doctor for further non-invasive evaluation.

The false negative rate was 2% meaning that the algorithm misses only 2% of abnormal images. This is much better than most of the labs. In our simulated cohort of 400

women, only 1 with abnormalities would be told that her results are normal. Given that this number is currently much higher, we believe this is a great result.

The true negative rate was 78%, meaning that the model was able to correctly identify 78% of images that did not actually have abnormalities.

Given the very low false negative rate of 2%, a 78% true negative rate is an amazing results. 282 out of 400 images could be classified as "No review required", leading to enormous costs savings.

5. Discussion and Conclusions

Our project achieved great results for all of the four experiments that we implemented. The accuracy for full image classification was lower when compared to cropped cells classification. This is probably because the cropped images contain the one cell only that has specific features that the model can learn.

For classifying the full images into five classes the best test accuracy of 80% was achieved by Resnet18. Classifying single cells achieved even better accuracy of 93% with the same model.

After putting the images into just two categories: normal and abnormal, we find Resnet50 achieves the best accuracy of 83%. Classifying single cells with a neural network worked even better and achieved the highest accuracy of 95% with Resnet18.

No deep learning model will have good results without good data. Good data are correctly labeled and unbiased (good coverage of all relevant kinds of data). The label for full images are loosely labeled as one class, where it might have more useful information. For five class images, some images might contain more than one class of cell, however the labels only conveys one class. For abnormal images, some images contain massive amounts of abnormal cells, however some contain only one, with the rest as background.

We expected that classification using two could lead to better results as opposed to five classes. When training on 5 classes, we have less samples for each class, whereas when we train on two classes, we have more examples for the model to learn. We find that the test accuracy does not really increase much for full images, but it increases a little bit for cropped images on average.

There are some limitations that we faced in our project that could also be addressed further. First, some of the images got misclassified probably because the slide was slightly rotated and part of the image included background. You can see examples in Figure 5 and 6. The discussed background is displayed in yellow as these images is are the normalized images seen by the network. This issue could be further addressed by pre-processing images or implementing a more standardized system to take these. Another limitation is that the cells on the full images are not labeled and therefore we were not able to proceed with a segmentation. If we were able to do the segmentation we could count the number of abnormal cells on an image and this could probably lead to even better results. A third limitation is the size of the data set. The more data, the more examples the model has to learn from. We would like to see more such data sets in the future to achieve better results.

These results are overall better then the currently commercialized systems. Future work go research further on who to implement this software in clinic, and how to combine it with appropriate hardware.

It is possible to program the models so that the final output are probabilities instead of just the predicted classes. This makes it possible for the user to set a specific threshold that would optimize the false negative rate. The higher the false negative rate the higher the true positive rate. The higher the true positive rate, the greater the savings, as less images have to be evaluated by humans. In the clinical settings however, it is expected that false negative rate should be as low as possible.

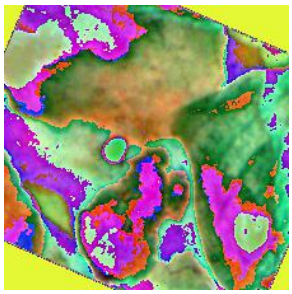

Figure 5. Abnormal classified as Normal (Full Image)

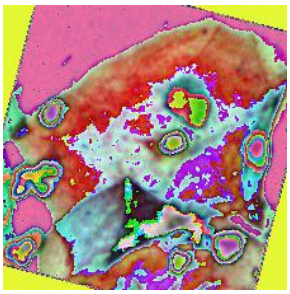

Figure 6. Normal classified as Abnormal (Full Image)

6. Individual Contributions

Both team members contributed to the project equally. All the tasks were done during meetings and the decision taken with approval of both sides. Wei Hung Hsu focused on fine tuning the model and Maja Garbulinska focused on the experiment design and result analysis.

References

[1] H. Ashtarian, E. Mirzabeigi, E. Mahmoodi, and M. Khezeli. Knowledge about cervical cancer and pap smear and the factors influencing the pap test screening among women. *International journal of community based nursing and midwifery*, 5(2):188, 2017.

[2] U. Banik, P. Bhattacharjee, S. U. Ahamad, and Z. Rahman. Pattern of epithelial cell abnormality in pap smear: A clinicopathological and demographic correlation. *Cytojournal*, 8, 2011.

[3] Y. Bar, I. Diamant, L. Wolf, and H. Greenspan. Deep learning with non-medical training used for chest pathology identification. In *Medical Imaging 2015: Computer-Aided Diagnosis*, volume 9414, page 94140V. International Society for Optics and Photonics, 2015.

[4] E. Bengtsson and P. Malm. Screening for cervical cancer using automated analysis of pap-smears. *Computational and mathematical methods in medicine*, 2014, 2014.

[5] R. Deepak, R. Kumar, N. Byju, P. Sharathkumar, C. Pournami, et al. Computer assisted pap smear analyser for cervical cancer screening using quantitative microscopy. *J Cytol Histol S*, 3:20–23, 2015.

[6] S. Kane, B. Khatibi, and D. Reddy. Higher incidence of abnormal pap smears in women with inflammatory bowel disease. *The American journal of gastroenterology*, 103(3):631, 2008.

[7] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. 2017.

[8] A. S. B. Reddy and D. S. Juliet. Transfer learning with resnet-50 for malaria cell-image classification. In *2019 International Conference on Communication and Signal Processing (ICCSP)*, pages 0945–0949. IEEE, 2019.

[9] J. Shi, R. Wang, Y. Zheng, Z. Jiang, and L. Yu. Graph convolutional networks for cervical cell classification. 2019.

[10] M. D. Stein, J. H. T. Fregnani, C. Scapulatempo, A. Mafra, N. Campacci, A. Longatto-Filho, and R. S. T. F. B. C. Hospital. Performance and reproducibility of gynecologic cytology interpretation using the focalpoint system: results of the rodeo study team. *American journal of clinical pathology*, 140(4):567–571, 2013.

[11] S. Vatathanavaro, S. Tungjitnob, and K. Pasupa. White blood cell classification: A comparison between vgg-16 and resnet-50 models.

[12] W. William, A. Ware, A. H. Basaza-Ejiri, and J. Obungoloch. A pap-smear analysis tool (pat) for detection of cervical cancer from pap-smear images. *Biomedical engineering online*, 18(1):16, 2019.