Performance of A Convolutional Neural Network in Screening Liquid Based Cervical Cytology Smears

Parikshit Sanyal, Sanghita Barui, Prabal Deb, Harish Chander Sharma
Departments of Pathology and Gynaecology and Obstetrics, Military Hospital Jalandhar Cantt, Punjab, Department of Pathology, Command Hospital, Alipore, Kolkata, West Bengal, India

Abstract

Context: Cervical cancer is the second most common cancer in women. The liquid based cervical cytology (LBCC) is a useful tool of choice for screening cervical cancer. Aims: To train a convolutional neural network (CNN) to identify abnormal foci from LBCC smears. Settings and Design: We have chosen retrospective study design from archived smears of patients undergoing screening from cervical cancer by LBCC smears. Materials and Methods: 2816 images, each of 256 × 256 pixels, were prepared from microphotographs of these LBCC smears, which included 816 “abnormal” foci (low grade or high grade squamous intraepithelial lesion) and 2000 ‘normal’ foci (benign epithelial cells and reactive changes). The images were split into three sets, Training, Testing, and Evaluation. A convolutional neural network (CNN) was developed with the python programming language. The CNN was trained with the Training dataset; performance was assayed concurrently with the Testing dataset. Two CNN models were developed, after 20 and 10 epochs of training, respectively. The models were then run on the Evaluation dataset. Statistical Analysis Used: A contingency table was prepared from the original image labels and the labels predicted by the CNN. Results: Combined assessment of both models yielded a sensitivity of 95.63% in detecting abnormal foci, with 79.85% specificity. The negative predictive value was high (99.19%), suggesting potential utility in screening. False positives due to overlapping cells, neutrophils, and debris was the principal difficulty met during evaluation. Conclusions: The CNN shows promise as a screening tool; however, for its use in confirmatory diagnosis, further training with a more diverse dataset will be required.

Keywords: Artificial intelligence, cervical cytology, liquid based smears, neural network, screening

INTRODUCTION

Cancer of the cervix uteri is the second most common cancer among women worldwide.[1] In India the incidence rate of cervical cancer is 14.7/100000 women per year, making it the second most common cancer in Indian women.[2] Liquid based cervical cytology (LBCC) has emerged as the standard of care for screening cervical cancer, due to its higher sensitivity and specificity than conventional smears, reduced artifacts such as blood, mucus, and other debris in the slide, reduced the rate of false negative diagnoses and a reduction in inadequate smears.[3-5] Screening LBCC for abnormal cells is a demanding and labor intensive task requiring a significant investment in man hours.[6] Thus, there has been interest in automated screening of cervical smears.[7] Presently, the Focal Point Slide Profiler (FSPS) ® (BD Tripath Imaging) and the ThinPrep Imaging System ® have been approved by FDA as primary screening tools for cervical cytology smears.[8-12] However, both of these systems are closed source, with proprietary rights and are tightly adherent to their respective devices, and require staining methods directed by the company. There is a need for a generic image analysis tool for screening liquid based cervical cytology (LBCC) smears.

Artificial neural networks (ANN) are systems of linear algebra that mimic the way the brain computes information. It calibrates it’s coefficients as to perform a certain task, e.g., pattern recognition. Thus, it has the ability to build up its own rules, referred to as “experience”.[13] Convolutional neural networks (CNNs) are a special class of ANNs which take a whole image as input and classify the image in defined categories. The input image is passed through multiple “layers”
in a feed forward manner, each layer comprising multiple, independent, and linear convolutional filters. Each layer passes its output to the next layer, with an overlaid nonlinearity. The network is trained by a method of back propagation, i.e. adjusting the coefficients of the linear equations in each layer, until the desired output is achieved. Further details about CNNs, the significance of each of their components, and how they perform image classification have been described by Karpathy et al.\[15\]

In the present study, we have chosen a CNN to identify foci of abnormality from LBCC smears. The smears were first classified as per the Bethesda System 2014, and then categorised in two broad strata:\[16\]
1. “normal” - comprising of the broad group “Negative for intraepithelial lesion or malignancy” (NILM), including normal epithelial cells, reactive, and inflammatory changes, infections
2. “abnormal” - comprising of Low grade squamous intraepithelial lesions (LSIL) and High grade squamous intraepithelial lesions (HSIL)

The objective of the present study was to train the CNN to identify foci of HSIL and LSIL and report them as abnormal. Subsequently, the foci are displayed to the pathologist for further evaluation. Effectively, the CNN would perform the role of a slide screener for LBCC smears.

**Subjects and Methods**

Cases undergoing screening for cervical cancer by LBCC smears were selected from a tertiary care hospital. A total of 36 LBCC smears were prepared. The BD SurePath™ technique and instrument was used for slide preparation.\[17\] 89 images from “abnormal” smears (HSIL and LSIL) and 462 images from “normal” smears were microphotographed, using a Nikon microphotography system. The various categories of diagnoses are shown in Table 1.

Each image was of 1280 × 720 pixels resolution. Images were systematically sliced by the ImageMagick™ command line tool into 256 × 256 foci.\[18\] A single image produced multiple such foci. After slicing into foci, blurred foci with indistinct features were manually removed. A total 2816 foci, each of 256 × 256 resolution, were selected for training and evaluation. The entire set (N = 2816) was split into random sets as shown in Table 2. No duplication was allowed between the sets.

A CNN was developed with the Python language, using the Keras platform.\[19\] The training method published by Cholet et al. was adopted.\[20\] The architecture of the network is shown in Figure 1. The network takes an image of size 3 × 256 × 256 (the three channels red, green and blue are represented separately in a color image), applies successive convolution and pooling layers until an output of 0 or 1 is produced.

The CNN was trained on the training set twice:
1. once with 20 epochs (Model A)
2. once with 10 epochs (Model B).

Thus, two different learning models were prepared. In each epoch, 500 batches of 16 images each were randomly selected by the network for training. The network

---

**Table 1: Distribution of LBCC images by diagnostic category as per Bethesda System 2014 (n=551)**

| Category | “Normal” | “Abnormal” |
|----------|----------|------------|
| Diagnosis by Bethesda System | Normal cellular elements only OR Reactive cellular changes associated with inflammation | Shift in flora suggestive of bacterial vaginosis | HSIL | LSIL |
| Number of cases (smears) | 22 | 06 | 04 | 04 |
| Number of microphotographs | 392 | 70 | 58 | 31 |

Figure 1: Architecture of the Convolutional Neural Network
self-calibrated its parameters over the period of training. Concurrent testing was carried out during training in the “Testing” dataset to keep track of learning by the CNN, as seen in Figure 2.

The accuracy on Testing set gradually increased with training, as seen in Figure 3. In Model A (20 epochs), the accuracy peaked after 10 epochs and then settled at 92.25%; the loss function (error rate) stabilized at 0.4. In Model B (10 epochs), a higher accuracy (94.75%), and lower error rate (0.23) was observed. This might be attributed to the learning rate parameter of the CNN, which causes the accuracy to decrease after reaching a peak in middle of training. However, both the models were preserved for evaluation.

The trained models were then run on the Evaluation dataset. Results were statistically interpreted by the R software package.[21]

RESULTS

Concurrent testing during training yielded the following results [Table 3]. In Model A, 92.25% diagnostic accuracy was achieved after 20 epochs of training. 13 (6.5%) foci were falsely labeled as “abnormal” by the CNN [Figure 4]. In Model B, the false positive rate was slightly higher (7%) during training, but the accuracy was also higher (94.75%), due to the reduced false negative rate (7 out of 200 foci, 3.5%).

After completion of training, both the models were run on the Evaluation set [Figure 5], which shows the following results [Table 4].

Table 2: Splitting the image data in training, testing and evaluation categories (n=2816)

|        | Abnormal | Normal | Total |
|--------|----------|--------|-------|
| Training | 410      | 410    | 820   |
| Testing  | 200      | 200    | 400   |
| Evaluation | 206    | 1390   | 1596  |
| Total    | 816      | 2000   | 2816  |

Table 3: Results on concurrent testing dataset (n=400)

|        | Model A | Model B |
|--------|---------|---------|
| Normal | Abnormal | Normal | Abnormal |
| Labeled by CNN | | | |
| Normal | 187     | 18   | 186 | 7 |
| Abnormal | 13     | 182  | 14  | 193 |
| Total   | 200     | 200  | 200 | 200 |

Figure 2: Training of the CNN (Model A)

Figure 3: Accuracy and loss function of two models plotted against epochs of training
Among 1390 images of the ‘normal’ class, 169 images (12.2%) were falsely labeled “abnormal” by Model A. Also, 31 out of 206 images of the “abnormal” class (15%) were missed by the CNN and falsely labeled “normal”.

Model B showed higher sensitivity (95.10%) and is thus suitable for screening; however, specificity was lower than Model A [Table 5]. A combination analysis of both models (i.e. a focus must be labeled “abnormal” by either of the two models to be diagnosed abnormal) showed 95.63% sensitivity [Table 6]. 197 out of 206 abnormal foci were correctly labeled “abnormal” by the combined model.

**DISCUSSION**

Most of the approaches to automated image analysis have focused on cell segmentation, which has remained an elusive problem. Early approaches to segmentation included geometric image analysis techniques such as mean-shift, median filtering, adaptive thresholding, Canny edge detection, edge detection by Riemannian dilation, and Hough transform for finding candidate nuclei. Doudkine et al. approached the segmentation problem by analyzing texture features from slides stained with quantitative stains for DNA. They used descriptive statistics of chromatin distribution, discrete texture features, ranges, Markovian, run length and fractal texture features for image classification. More recently, Rodenacker et al. prepared a set of parameters for extracting features from cytological images. A geometrical segmentation approach was also used by Anderson et al., who reported a sensitivity of 95% for severe dysplasia and 90% for moderate dysplasia.

A combination of nuclear texture features was found which could reliably classify the images. Good segmentation of LBCC smears was achieved by Zhang et al.; their segmentation method achieved 93% accuracy for cytoplasm, and 87.3% F-measure for nuclei.

Neural networks provide an alternative approach to the problem. These networks process an image in entirety and produce an output. After repeated epochs of reinforcement...
training, the network adjusts its parameters to produce the correct result in majority of cases. Thus, the segmentation problem is bypassed. A very early neural network was the PAPNET system, developed in the 1990s. Over the last two decades, the convolutional neural network (CNN) model has proved to be a reliable image classifier in several scenarios, including recognizing everyday objects, traffic signs, text, and handwritten numbers. The CNN model has been chosen for the purpose because of its consistently superior performance than other machine learning models.

The CNN extracts features from an image in its successively deeper layers [Figure 6], until the image is converted to a single number “0”, which corresponds to “abnormal”, or “1”, corresponding to “normal”.

The principal difficulties met in this study was that of false positives, unlike Anderson et al. 280 out of 1390 normal foci (20%) were marked “abnormal” by the combined model. The false positives observed in this study might be attributable to overfitting, i.e. training to both the “signal” (i.e. abnormal cells) and “noise” (artifacts) in the Training data. In a few cases, overlapping cells produced hyperchromasias in the image, which has been wrongly labeled as “abnormal” by the CNN [Figure 7]. Also, we have randomly sliced images, so that neutrophils, background debris and hemorrhage have all been included in the evaluation set. A few of these foci have been falsely marked positive [Figure 4], indicating the need for further training.

The high sensitivity of the CNN in picking up abnormal foci (95.63%) makes it suitable for screening. To improve the specificity so that the CNN becomes useful for diagnostic purposes, further training with a larger dataset will be required. Presently, the CNN is not useful for the purpose of confirmatory diagnosis, because of very low positive predictive value. However, the high negative predictive value (99.19%) indicates its potential for use in screening. With an automated slide preparation system, microphotography and slide scanning system, the CNN can provide reproducible results and become a useful screening tool.

**Conclusion**

The present study demonstrates the performance characteristics of a convolutional neural network in screening liquid based cervical smears. Training with a larger and more diverse dataset will be required before it can be employed for confirmatory diagnosis.

**Financial support and sponsorship**

Nil.

**Conflicts of interest**

There are no conflicts of interest.

**References**

1. International Agency for Research on Cancer (IARC) Information Centre on HPV and Cancer. Available from: http://www.hpvcentre.net/statistics.php. [Last accessed on 2018 Dec 12].
2. Swaminathan S, Katoch VM (eds). Consolidated Report of Hospital Based Cancer Registries 2012-2014. Bengaluru: Indian Council of Medical Research; 2016.

3. Abulafia O, Pezzullo JC, Sherer DM. Performance of ThinPrep liquid-based cervical cytology in comparison with conventionally prepared Papanicolaou smears: A quantitative survey. Gynecologic Oncol 2003;90:137-44.

4. Bernstein SJ, Sanchez-Ramos L, Nqubisi B. Liquid-based cervical cytologic smear study and conventional Papanicolaou smears: A metaanalysis of prospective studies comparing cytologic diagnosis and sample adequacy. Am J Obstet Gynecol 2001;185:308-17.

5. Nanda K, McCrory DC, Myers ER, Bastian LA, Hasselblad V, Hickey JD, et al. Accuracy of the Papanicolaou test in screening for and follow-up of cervical cytologic abnormalities: A systematic review. Ann Intern Med 2000;132:810-9.

6. Elsheikh TM, Austin RM, Chhieng DF, Miller FS, Moriarty AT, Renshaw AA. American society of cytopathology workload recommendations for automated pap test screening: Developed by the productivity and quality assurance in the era of automated screening task force. Diagn Cytopathol 2013;41:174-8.

7. Bengtsson E, Malm P. Screening for Cervical Cancer Using Automated Analysis of PAP-Smears. Computational and Mathematical Methods in Medicine 2014:2014:12. doi: 10.1155/2014/842037.

8. Kardos TF. The FocalPoint System: FocalPoint slide profiler and FocalPoint GS. Cancer 2004;102:334-9.

9. Quddus MR, Neves T, Reilly ME, Moriarty AT, Renshaw AA. American society of cytopathology workload recommendations for automated pap test screening: Developed by the productivity and quality assurance in the era of automated screening task force. Diagn Cytopathol 2013;41:174-8.

10. Linder J, Zahniser D. The ThinPrep Pap test. A review of clinical studies. Acta Cytol 1997;41:30-8.

11. Monsonego J, Autillo-Touati A, Bergeron C, Dachez R, Liaras J, Saurel J, et al. Liquid-based cytology for primary cervical cancer screening: A multi-centre study. Br J Cancer 2001;84:360-6.

12. Wilbur DC, Bibbo M. Automation in cervical cytology. In: Comprehensive Cytopathology (3e). New York: Elsevier; 2008. p. 1021.

13. Haykin S. Neural Networks: A Comprehensive Foundation (2e). New York: Prentice Hall; 1999.

14. Krizhevsky A, Sutskever I, Hinton G. ImageNet Classification with Deep Convolutional Neural Networks. Neural Information Processing Systems. Available from: https://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf. [Last accessed on 2018 Dec 12]

15. Karpathy A. Stanford University Open Access Course. Convolutional Neural Networks in for Visual Recognition. Available from: http://www.cs231n.github.io/convolutional-networks/. [Last accessed on 2018 Sep 17]

16. Nayar R, Solomon D. The Bethesda System for reporting cervical cytology atlas, website, and Bethesda interobserver reproducibility project. Cytojournal 2004;1:4.

17. BD SurePath Liquid Based Pap Test. Available from: https://www.bd.com/en-us/offering/cervical-cancer-screening/cervical-sample-collection/surepath-liquid-based-pap-test. [Last accessed on 2018 Dec 20]