Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/mjafi

Original Article

Performance characteristics of an artificial intelligence based on convolutional neural network for screening conventional Papanicolaou-stained cervical smears

Parikshit Sanyal ^a, Prosenjit Ganguli ^{b,*}, Sanghita Barui ^a^a Graded Specialist (Pathology), Military Hospital Jalandhar Cantt, India^b Professor & Senior Adviser (Pathology), Command Hospital (Eastern Command), Alipore, Kolkata 27, India

ARTICLE INFO

Article history:

Received 15 January 2019

Accepted 6 August 2019

Available online 11 December 2019

Keywords:

Cervical cancer

Screening

Papanicolaou smear

Machine learning

Neural network

Artificial intelligence

ABSTRACT

Background: The conventional Papanicolaou-stained cervical smear is the most common screening test for cervical cancer. The sensitivity of the test in detecting abnormal cells is 67–75% in various studies. Owing to the volume of smears at cancer screening centres, significant man-hours are expended in the test. We have developed a software program for identification of foci of abnormal cells from conventional smears. We have chosen the convolutional neural network (CNN) model for its efficacy in image classification.

Methods: A total of 1838 microphotographs from cervical smears, containing 1301 'normal' foci and 537 'abnormal' foci were included in the study. The data set was split into training, testing and validation sets. A CNN was developed in the Python programming language. The CNN was trained with the training and testing set. At the end of training, 94.64% accuracy was achieved in the testing set. The CNN was then run on the validation set (441 images).

Results: The CNN showed 94.28% sensitivity, 96.01% specificity, 91.66% positive predictive value and 97.30% negative predictive value. The CNN could recognise normal squamous cells, overlapping cells, neutrophils and debris and classify the focus appropriately. False positives were reported when the CNN failed to recognise overlapping cells (2.7% microphotographs). It could correctly label cell clusters with high nuclear cytoplasmic ratio and hyperchromasia. In 1.8% of microphotographs, a false negative was reported.

Conclusion: The CNN showed 95.46% diagnostic accuracy, suggesting potential use in screening.

© 2019 Director General, Armed Forces Medical Services. Published by Elsevier, a division of RELX India Pvt. Ltd. All rights reserved.

* Corresponding author.

E-mail address: prosenjitganguli@hotmail.com (P. Ganguli).<https://doi.org/10.1016/j.mjafi.2019.08.001>

0377-1237/© 2019 Director General, Armed Forces Medical Services. Published by Elsevier, a division of RELX India Pvt. Ltd. All rights reserved.

Introduction

Cervical cancer is the second most common cancer among Indian women.¹ The incidence of cervical cancer in India was reported to be 14.7 to 24.3 per 100,000 women in various studies.^{2,1} Majority of cancer screening centres in India still routinely use the Papanicolaou-stained conventional cervical smear, which is prepared manually by making a direct smear from the cervix, using a cytobrush or spatula. Although liquid-based cytology has emerged as the standard of care in screening cervical cancer,^{3,4} most of the centres who conduct cervical cancer screening do not have facilities for liquid-based cytology. Manual screening of cervical smears by microscopy is a labour-intensive task requiring thorough scrutiny by a cytotechnician or pathologist. Automated screening of conventional cervical smears will significantly reduce man-hours invested in screening cervical cancer. An effective screening system will mark particular foci in a given smear which have some abnormality and present them to the pathologist for further evaluation. Such screening systems have been developed for liquid-based cytology;^{5,6} however, there are no such solutions for conventional cervical smears.

Artificial neural network (ANN) is a machine learning model that emulates the mechanism of learning by a human brain. A 'neuron' in an ANN is defined as a mathematical function, which, when given a number as input, can apply the function and pass the result to another neuron. The functions used by a neuron may be sigmoid function, rectified linear unit (ReLU) function or inverse tangent (tanh) functions.⁷ A variant of ANN the convolutional neural network (CNN) uses a special function called 'convolution', which is a mathematical overlap of two matrices. The architecture is similar to the ventral visual pathway of the brain, where the layers represent a particular retinotopic area of the brain.^{8,9} The image features extracted by the layers are finally fed into a classifier that determines the category the image belongs to. The mathematical formalism and working principle of a CNN has been described in detail by Karpathy.¹⁰ CNNs have been successful in recognising objects from images and categorising images into classes.¹¹ A CNN can be 'trained' with a set of example images. For a binary image classifier, two sets of images, one 'abnormal' and another 'normal', are shown to the CNN. The CNN calibrates its parameters so as to produce the correct classification on this example set of images. Training is imparted over several 'epochs'; at each epoch, the CNN reads the entire data set and tries to predict the correct classification for each datum. After each epoch, the performance of the CNN is measured with a 'test' data set. After each epoch of training, the accuracy of the network should improve; the error rate or 'loss' should reduce with training.

We chose a CNN machine learning model to classify conventional cervical smears because of the well-demonstrated performance of this model in image classification. The objective of our study was to develop a CNN which will identify 'abnormal' foci in conventional cervical smears (i.e., foci of cells with high N:C ratio, nuclear hyperchromasia and irregular nuclear membrane), for further evaluation by a pathologist. We have chosen to include all random foci from microphotographs of conventional cervical smears,

containing squamous epithelial cells, endocervical cells, overlapping cell clusters, debris and neutrophils, i.e., usual constituents of the conventional smear. Our purpose was to test whether a CNN, with adequate training, could identify abnormal foci even from this 'noisy' data set, a task that till date could only be performed by a human. Owing to the nature of the data set, the problem at hand is not simply that of image analysis but a nontrivial machine learning problem.

Materials and methods

Conventional cervical smears were collected from the archives of a tertiary care hospital and a cervical cancer screening centre of the East India. Fourteen smears of category 'normal' (including Negative for Intraepithelial Lesion or Malignancy and Reactive Changes associated with Inflammation) were selected for the study. In addition, 4 cases of 'abnormal' smears (2 low-grade squamous intraepithelial lesions and 2 high-grade squamous intraepithelial lesions) were selected. Microscopic foci from these smears were selected by a group of pathologists to be representative of the entire smear. An Olympus integrated microphotography system was used to photograph the slides at 40 $\times$ resolution. Two hundred twenty-two microphotographs were taken, each of resolution 2560 $\times$ 1920 pixels and 8 bits of colour space. The microphotographs were taken in the same illumination and magnification of the microscope and after correction of white balance. The microphotographs included 58 foci marked as 'abnormal' by the pathologist and 164 foci with no abnormalities ('normal'). The microphotographs were then systematically cropped using a computer software program (ImageMagick) into smaller images of 512 $\times$ 512 pixels each, to be used for the CNN. A total of 1838 images were produced.

This data set was randomly segregated in two subsets. A 'training' subset of 1397 images was prepared for training the CNN. The remaining 441 images were kept for performance evaluation ('validation') of the CNN after completion of training. Distribution of the images in various data sets is shown in Table 1. Of the 537 images in the 'abnormal' category, 212 images also show reactive changes associated with inflammation. However, the images were selected from reported cases of atypical squamous cells of undetermined significance (ASC-US)/low-grade squamous intraepithelial lesion (LSIL)/high-grade squamous intraepithelial lesion (HSIL), and consensus diagnosis on these images was reached before inclusion.

The CNN was developed in the Python language, using the OpenCV, TensorFlow and Keras libraries, in the Google Colab Research platform. Because of restrictions of memory usage in the Colab platform, using images of 512 $\times$ 512 pixels directly

Table 1 – Splitting of microphotographs in different data sets (N = 1838).

Category	Training	Validation	Total
Abnormal	397	140	537
Normal	1000	301	1301
Total	1397	441	1838

was not possible. Thus, they were resized to images of 128×128 pixels so that they could be used as inputs to the model. The architecture of the CNN is shown in Fig. 1. The CNN accepts an image of size $128 \times 128 \times 3$ (the three colour channels red, green and blue are represented separately in colour images) and produces an output '0' (abnormal) or '1' (normal).

The CNN was then trained with examples from the training set. Of the 1307 images in the training set, 80% of images were used as labelled examples for learning. The rest 20% were used for concurrent testing of performance after each epoch. A training session is shown in Fig. 2.

The network was trained with a batch size of 64, for 40 epochs. After the end of training, 97% accuracy was achieved in predicting the labels of the testing data set. The accuracy and error ('loss') of the CNN over epochs of training is shown in Fig. 3.

After training, the CNN was run on the validation data set. Results were interpreted by standard statistical methods. Because cervical smear is a screening test, finding a single

'abnormal' focus in an entire slide will mark the whole slide as 'abnormal' and it will be put up for review by the pathologist. Thus, we have compared the sensitivity of detecting single focus of 'abnormal' cells with whole slides. However, because finding a 'normal' focus in a slide does not ensure that all foci from the slide will be classified as 'normal'. Thus, specificity of analysis by a single focus and that of the whole slide are not comparable.

Results

The performance characteristics of the CNN are shown in Table 2.

Sensitivity of the CNN in recognition of abnormal foci was found to be 94.28% and specificity, 96.01%. The positive predictive value (PPV) was found to be 91.66%, and the negative predictive value (NPV) was 97.30%.

Twelve foci from 5 cases were reported falsely positive in the present study (2.7%), when overlapping cell clusters were mistaken for hyperchromasia (Fig. 4f). Four of these cases were diagnosed to be negative for intraepithelial lesion or malignancy (Bethesda system), and one case was diagnosed to have reactive changes associated with inflammation.

Fig. 4 shows a subset of validation data set with the predictions made by the CNN. The CNN recognised normal superficial cells (Fig. 4a), overlapping cells (Fig. 4b), cells infiltrated with neutrophils (Fig. 4c), slightly dense stained cells (Fig. 4d), neutrophils and debris in background (Fig. 4e). In a few cases, it has failed to recognise cellular overlap and variations in staining intensity, resulting in false positive (Fig. 4f). It has also correctly labelled clusters with high N:C ratio and hyperchromasia (Fig. 4g), identified isolated cells with high N:C ratio (Fig. 4h and k), recognised overlapping cells in dense clusters (Fig. 4j) and scattered cells with hyperchromatic nuclei (Fig. 4m). A few false negatives were also met (Fig. 4l).

Discussion

Image analysis tools are usually based on geometric functions, such as edge detection, Gaussian blur, watershed transform, etc. Such methods are useful for analysis of images with well-segmented cells without background debris or overlap. However, when analysing a noisy data set such as conventional cervical smears, they are often inadequate. The model developed in this study does not extract any specific parameters from an image, which is the point of difference of our approach from a conventional image analysis approach. In conventional image analysis modules, features of the image such as nuclear cytoplasmic ratio, nuclear membrane irregularity and clustering of nuclei are extracted. The data extracted from these features are then used as input to a suitable machine learning tool (i.e., logistic regression) to arrive at a final prediction.¹² We have used a CNN, which does not extract specific parameters of the image but takes the whole image (i.e., numerical data of each pixel of the image) as an input (a three-dimensional array of numbers) and converts it to a single number output, '0' or '1' (Fig. 5).

Fig. 1 – Architecture of the CNN. CNN, convolutional neural network.

```

Epoch 29/40
1117/1117 [=====] - 5s 5ms/step - loss: 0.0073 - acc: 0.9991 - val_loss: 0.4301 - val_acc: 0.9607
Epoch 30/40
1117/1117 [=====] - 5s 5ms/step - loss: 0.0072 - acc: 0.9991 - val_loss: 0.4652 - val_acc: 0.9571
Epoch 31/40
1117/1117 [=====] - 5s 5ms/step - loss: 0.0054 - acc: 0.9991 - val_loss: 0.4423 - val_acc: 0.9536
Epoch 32/40
1117/1117 [=====] - 5s 5ms/step - loss: 0.0049 - acc: 0.9991 - val_loss: 0.4253 - val_acc: 0.9500
Epoch 33/40
1117/1117 [=====] - 5s 5ms/step - loss: 0.0056 - acc: 0.9991 - val_loss: 0.4742 - val_acc: 0.9464
Epoch 34/40
1117/1117 [=====] - 5s 5ms/step - loss: 0.0047 - acc: 0.9991 - val_loss: 0.4876 - val_acc: 0.9536
Epoch 35/40
1117/1117 [=====] - 5s 5ms/step - loss: 0.0047 - acc: 0.9991 - val_loss: 0.4579 - val_acc: 0.9536
Epoch 36/40
1117/1117 [=====] - 5s 5ms/step - loss: 0.0057 - acc: 0.9991 - val_loss: 0.4870 - val_acc: 0.9500
Epoch 37/40
1117/1117 [=====] - 5s 5ms/step - loss: 0.0047 - acc: 0.9991 - val_loss: 0.4922 - val_acc: 0.9464
Epoch 38/40
1117/1117 [=====] - 5s 5ms/step - loss: 0.0041 - acc: 0.9991 - val_loss: 0.4803 - val_acc: 0.9500
Epoch 39/40
1117/1117 [=====] - 5s 5ms/step - loss: 0.0053 - acc: 0.9991 - val_loss: 0.5192 - val_acc: 0.9393
Epoch 40/40
1117/1117 [=====] - 5s 5ms/step - loss: 0.0034 - acc: 0.9991 - val_loss: 0.5096 - val_acc: 0.9464

```

Fig. 2 – Training session showing accuracy of 94.64%.

Fig. 3 – Accuracy and error rate (loss) of the CNN over epochs of training. CNN, convolutional neural network.

Table 2 – Contingency table of performance of the CNN over the validation set (N = 441).

Parameter	Predicted label	True label		Total
		Abnormal	Normal	
Predicted label	Abnormal	132	12	144
		True positive (TP)	False positive (FP)	
	Normal	8	289	297
		False negative (FN)	True negative (TN)	
	Total	140	301	441
Sensitivity	TP/(TP + FN)	132/140	94.28%	
Specificity	TN/(FP + TN)	289/301	96.01%	
Positive predictive value	TP/(TP + FP)	132/144	91.66%	
Negative predictive value	TN/(TN + FN)	297/301	97.30%	
Diagnostic accuracy	(TP + TN)/Total	421/441	95.46%	

CNN, convolutional neural network.

Fig. 4 – Images from the validation set with their original label and predicted label (Pap stain, 40×). (a) Normal squamous cells; (b) Normal intermediate cell; (c) Overlapping cells; (d) Normal squamous cells with inflammation and artifacts; (e) Normal squamous cells with debris and lactobacilli; (f) Overlapping squamous cells; (g) Squamous cells with high N:C ratio; (h) Squamous cells with high N:C ratio and nuclear hyperchromasia; (j) Squamous cells in cluster with high N:C ratio; (k,l) Isolated squamous cells with high N:C ratio; (m) Dispersed squamous cells with high N:C ratio.

The initial problem is that of cell segmentation, i.e., recognising individual cells from overlapping clusters. Wang et al.¹³ have achieved high effective segmentation on 362 images of cervical smears (sensitivity: $94.25\% \pm 1.03\%$ and

specificity $93.45\% \pm 1.14\%$) by mean shift clustering. Their method of identifying dysplasia is reported to have 96% accuracy.¹³ However, they operated on extracted features from the image and not on the whole image. The present study

Fig. 5 – Intermediate layers of the CNN. CNN, convolutional neural network.

uses a CNN that does not rely on segmentation but operates on the whole image at once. The diagnostic accuracy achieved by the CNN (95.46%) is comparable to that achieved by Wang et al.¹³

The second problem is that of overlay artifacts, i.e., neutrophils and debris, which interfere with cellular features. Two approaches have been used in previous studies to overcome this problem.

- (a) A conventional image analysis module can be used to extract specific features from the image, such as nuclear cytoplasmic ratio, nuclear membrane irregularities, clustering of nuclei, and these data can be used as input for a neural network. In this approach, the whole image is not passed to the neural network, thus reducing noise from overlay artifacts. Such an approach has been used by Savala et al.¹⁴ in the analysis of thyroid cytopathology smears.
- (b) Filters might be used on the image, such as watershed transform, blurring and thresholding, so that only the cellular elements are read by the neural network.

In the present study, no attempt was made to remove overlay artifacts and debris. Whole images, including debris, were used as input to the neural network. It has been mentioned in literature¹¹ that neural networks can learn to distinguish between ‘signal’ (cellular features) and ‘noise’ (debris) over repeated epochs of training. This feature of neural networks has been used in this study. It is possible that some features such as small cells were lost by the neural network, but it is not possible to fully visualise the middle layers of a CNN, and thus, information lost cannot be exactly quantified.

Whether any smaller cell/bad nuclei were lost by the neural network is not known because it is not possible to fully visualise the middle layers of a CNN.

There have been two notable studies regarding screening conventional cervical smears with closed source, proprietary slide screening systems. Saieg et al.¹⁵ had examined 120 conventional cervical smears with the BD Focal Point Imaging System® and reported sensitivity and specificity to be 100% and 70.3%, respectively. We did achieve less sensitivity (94.28%) than Saieg et al.¹⁵ Stevens et al.¹⁶ used the Autopap 300® in classifying 1840 conventional smears and reported 82.7% sensitivity. However, they focused on recognising false-negative smears only and did not include previously diagnosed ‘abnormal’ smears in the study. In contrast, the present study included smears from both categories, ‘abnormal’ and ‘normal’, and both true- and false-positive rates are reported in the study. The sensitivity in the present study (94.28%) was better than that of the study by Stevens et al.¹⁶

As per our knowledge, the present study is the only one that uses the whole microphotographic focus as an input to a CNN, to automatically screen conventional cervical smears. The CNN does not depend on extracted features from the image but can generate predictions from whole images of microphotographs from conventional cervical smears.

The difficulties encountered in analysing images from conventional smears are due to overlapping cells, background debris and neutrophils. The present study demonstrates that a CNN, with training over many epochs and example data sets, can learn to recognise these features, much similar to the trainee pathologist.

The working mechanism of the CNN is shown in Fig. 5. A 128×128 colour image is convoluted in several layers, each time simplified than the previous layer. The brighter spots in the intermediate layers correspond to the features of the image crucial for the final classification. A final flattened layer is produced, before a binary output ‘0’ or ‘1’.

Twelve foci from five cases were reported falsely positive in the present study, when overlapping cell clusters were mistaken for hyperchromasia (Fig. 4f). Four of these cases were diagnosed to be Negative for Intraepithelial Lesion or Malignancy (Bethesda system), and one case was diagnosed to have Reactive Changes Associated with Inflammation. In a few focus, endocervical cells were mistaken for atypical cells. In a few focus where dysplastic cells were present among numerous normal squamous cells, the CNN generated a false-negative report.

The domain of the CNN is limited only to the conventional cervical smear. The CNN is not a generalised cell classifier. It has been only trained with a specific data set of conventional smears. However, given any random image of any object, it will produce an output of ‘0’ or ‘1’. It is important that the CNN be applied in its proper context. A tertiary care centre with an automated microscope equipped with the CNN could serve as an automated screening terminal for pooled cervical smears from many different screening centres. The role of the CNN will be to identify foci of abnormality from smears and display them to the pathologist. It will remain for the pathologist to decide whether to accept the classification by the CNN or to reject it.

Conclusion

The present study demonstrates the ability of a CNN to identify foci of abnormality from conventional cervical smears. A large, multicentric study for assessment of its diagnostic accuracy will be required, before the CNN can be

used as a screening terminal for conventional cervical smears.

Conflicts of interest

The authors have none to declare.

REFERENCES

- Swaminathan SKV. *Consolidated Report of Hospital Based Cancer Registries 2012-2014*. Bengaluru: Indian Council of Medical Research; 2016.
- Sreedevi A, Javed R, Dinesh A. Epidemiology of cervical cancer with special focus on India. *Int J Womens Health*. 2015;7:405–414.
- Abulafia O, Pezzullo JC, Sherer DM. Performance of ThinPrep liquid-based cervical cytology in comparison with conventionally prepared Papanicolaou smears: a quantitative survey. *Gynecol Oncol*. 2003;90:137–144.
- Lee KR, Ashfaq R, Birdsong GG, Corkill ME, McIntosh KM, Inhorn SL. Comparison of conventional Papanicolaou smears and a fluid-based, thin-layer system for cervical cancer screening. *Obstet Gynecol*. 1997;90:278–284.
- Cervical Cancer Screening - BD Internet. Cited 2019 Jan 10; Available from <https://www.bd.com/en-us/offerings/capabilities/cervical-cancer-screening>.
- Hologic diagnostic solutions: aptima and ThinPrep | HealthDxS internet. Cited 2019 Jan 10; Available from <https://healthdxs.com/en/>.
- Haykin S. *Neural Networks: A Comprehensive Foundation*. 2nd ed. New York: Prentice Hall; 1999.
- Mishkin M, Ungerleider LG. Contribution of striate inputs to the visuospatial functions of parieto-preoccipital cortex in monkeys. *Behav Brain Res*. 1982;6:57–77.
- Yamins DLK, DiCarlo JJ. Using goal-driven deep learning models to understand sensory cortex. *Nat Neurosci*. 2016;19:356–365.
- Karpathy A. CS231n convolutional neural networks for visual recognition internet. Cited 2019 Jan 10; Available from <http://cs231n.github.io/convolutional-networks/>.
- Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks internet. In: *Proceedings of the 25th International Conference on Neural Information Processing Systems – Volume 1*. USA: Curran Associates Inc.; 2012 cited 2019 Jan 10. page 1097–1105. Available from: <http://dl.acm.org/citation.cfm?id=2999134.2999257>.
- Sanyal P, Ganguli P, Barui S, Deb P. Pilot study of an open-source image analysis software for automated screening of conventional cervical smears. *J Cytol*. 2018;35:71–74.
- Wang P, Wang L, Li Y, Song Q, Lv S, Hu X. Automatic cell nuclei segmentation and classification of cervical Pap smear images. *Biomed Signal Process Control*. 2019;48:93–103.
- Savala R, Dey P, Gupta N. Artificial neural network model to distinguish follicular adenoma from follicular carcinoma on fine needle aspiration of thyroid. *Diagn Cytopathol*. 2017;46.
- Saieg MA, Motta TH, Fodra ME, Scapulatempo C, Longatto-Filho A, Stiepcich MMA. Automated screening of conventional gynecological cytology smears: feasible and reliable. *Acta Cytol*. 2014;58:378–382.
- Stevens M, Milne A, Kerrie J. Effectiveness of automated cervical cytology rescreening using the AutoPapt 300 QC system. *Diagn Cytopathol*. 1997;16.