Deep Convolutional Neural Networks For Detecting Cellular Changes Due To Malignancy

Håkan Wieslander*
Department of IT
Uppsala University
h.wieslander@hotmail.com

Gustav Forslid*
Department of IT
Uppsala University
gustav@forslid.com

Ewert Bengtsson
Department of IT
Uppsala University
ewert.bengtsson@it.uu.se

Carolina Wählby
Department of IT
Uppsala University
carolina.wahlby@it.uu.se

Jan-Michaél Hirsch
Department of Surgical Sciences
Uppsala University
jan.michael.hirsch@akademiska.se

Christina Runow Stark
Swedish Dental Service
Medical Dental Care, Södersjukhuset
christina.runow-stark@sll.se

Sajith Kecheril Sadanandan
Department of IT
Uppsala University
sajith.ks@it.uu.se

Abstract

Discovering cancer at an early stage is an effective way to increase the chance of survival. However, since most screening processes are done manually it is time inefficient and thus a costly process. One way of automizing the screening process could be to classify cells using Convolutional Neural Networks. Convolutional Neural Networks have been proven to be accurate for image classification tasks. Two datasets containing oral cells and two datasets containing cervical cells were used. For the cervical cancer dataset the cells were classified by medical experts as normal or abnormal. For the oral cell dataset we only used the diagnosis of the patient. All cells obtained from a patient with malignancy were thus considered malignant even though most of them looked normal. The performance was evaluated for two different network architectures, ResNet and VGG. For the oral datasets the accuracy varied between 78-82% correctly classified cells depending on the dataset and network. For the cervical datasets the accuracy varied between 84-86% correctly classified cells depending on the dataset and network. The results indicate a high potential for detecting abnormalities in oral cavity and in uterine cervix. ResNet was shown to be the preferable network, with a higher accuracy and a smaller standard deviation.

*Equal contribution

1. Introduction

Discovering cancer in an early stage leads to an early treatment, which lowers the risk of morbidity and mortality. By implementing a screening program the chances of discovering cancer early increases. This means more people with cancer can get an early treatment. For cervical cancer there has been a screening procedure based on the so called PAP-test available for around 75 years. The test is based on collecting cells by brushing the epthelial layer of the uterine cervix. According to WHO early treatment can prevent up to 80% of cervical cancer [19]. Oral cancer is in many ways similar to cervical cancer and the incidence is increasing world-wide. But there are no established screening programs for oral cancer today. This makes oral cancer a disease with high morbidity and mortality rates [20].

The conventional way of carrying out the cervical cancer screening is a visual examination under a microscope of the collected cells that have been smeared on a glass slide and stained. The cytotechnologist doing the screening looks for any cells showing signs of malignant changes flipping between low and high magnification to examine the typically around 100000 cells in a sample. This process can take around 10-15 minutes and is thus costly. There was therefor already in the 1950s attempts to automate this using automated image analysis. After numerous research projects around the world commercial image analysis based screening systems appeared around the turn of the millennium [1]. Unfortunately these systems use the same time-consuming

search for diagnostic cells and are not sufficiently cost effective to replace manual screening. The high cost leads to the fact that screening programs for cervical cancer are mostly performed in developed countries [18].

There is, however, an alternative way of doing the screening and that is to look for malignancy associated changes (MAC). MAC refers to small changes in the morphology and chromatin structure of the nucleus in a cell. The changes appear in normal looking cells located in the vicinity of tumor-associated areas [14]. The changes are too subtle to be caught by visual examination, but with computerized image analysis one can start to detect these changes. The advantage of looking for MAC is that the changes appear in basically all cells in the sample so instead of searching for a few clearly malignant cells we only need to do a careful analysis of the slight changes in the chromatin structure of a small random population of cells [9]. Conventional image analysis has been used in trying to develop screening systems based on the MAC approach but there is still need for improvements for this approach to reach a real breakthrough [10].

In recent years the deep learning field has exploded and found application in many different areas. Convolutional Neural Networks (CNN) are a type of deep learning networks. CNNs have been proven to produce high accuracy for image classification tasks and have won the ImageNet Large Scale Visual Recognition Challenge (ILSVCR) the last years [13]. CNN:s have also been applied to cervical cancer screening studies [2], [21], [6].

This paper presents a pilot study on applying the PAP-based screening method for early detection of oral cancer and to do so using the MAC approach and through application of CNN. As a baseline we also trained a CNN to recognize cervical cancer using the conventional approach of classifying diagnostic cells. We did this both on the publicly available Herlev dataset and on a larger and more difficult dataset collected in an earlier project at our center, where cells are not discarded due to debris. The oral dataset was collected for this work. Since we used the MAC approach we only needed the diagnosis on the patient level, not for the individual cells. To see how far the MAC effect could reach and to decrease the influence of clearly diagnostic cells in our dataset we also tried to detect cancer in samples collected from the opposite side of the mouth from where the malignant changes were seen.

2. Materials and Methods

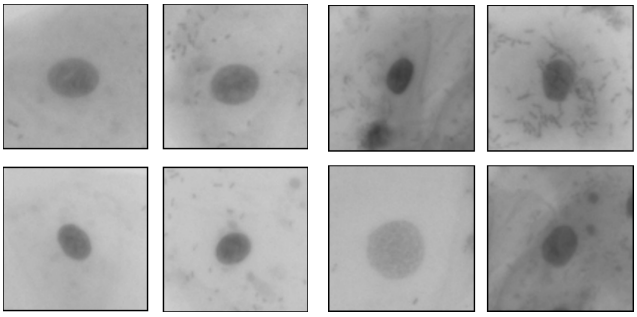
2.1. Datasets

2.1.1 Oral Dataset

The cell samples were collected at Södersjukhuset in Stockholm. The patients have mixed genders, are non smoking, some are human papillomavirus (HPV) positive and some are not, and they have an age span of 47-77 years. From each patient samples were collected with a brush that is scraped at areas of interest in the oral cavity. Each scrape is then smeared out on a glass, which is then stained to highlight important cellular structures. Images from each sample were acquired using an Olympus BX51 bright-field microscope with a 20x, 0.75 NA objective giving a pixel size of 0.32 μm. The microscope was equipped with an E-662 Piezo server controller and actuator which controls movement of the microscope objective so that the focus level changes slightly. From each smear, areas with a large number of cells were selected manually and a stack of images was acquired with a step length of 0.4 μm. Each stack contained 15 images with different focus, so that each cell was in focus in at least one of the images in the stack.

Table 1: Oral dataset. The table presents the diagnosis of each sample and how the dataset was divided for training and testing. (hs) means healthy side and lists samples collected from the healthy side of the oral cavity of patients with tumors.

				Oral Dataset 1		Oral Dataset 2	
Patient	Sample	Diagnosis	Nr. of cells	Training (Fold)	Testing (Fold)	Training (Fold)	Testing (Fold)
1	1	Healthy	2123	1,2,3	-	1,2	-
	2	Healthy	1993	1,2,3	-	1,2	-
2	1	Healthy	1454	1,2,3	-	1,2	-
	2	Healthy	1024	1,2,3	-	1,2	-
3	1	Healthy	928	-	1,2,3	-	1,2
	2	Healthy	777	-	1,2,3	-	1,2
4	1	Tumor	245	1,3	2	-	1,2
5	1	Tumor (hs)	1198	-	1,2,3	1	2
	2	Tumor (hs)	1098	-	1,2,3	1	2
	3	Tumor	519	2,3	1	-	1,2
6	1	Tumor	988	1,2	3	-	1,2
	2	Tumor	828	1,2	3	-	1,2
	3	Tumor (hs)	872	-	1,2,3	2	1
	4	Tumor (hs)	912	-	1,2,3	2	1



(a) Cells from a healthy patient. (b) Cells from a patient with tumor.

Figure 1: Example images from the oral dataset

The dataset was divided into two sub datasets. The first one (Oral Dataset 1) containing samples taken from healthy patients and samples taken from the tumor side of the oral cavity of patients with tumors. The second (Oral Dataset 2) contained samples taken from healthy patients and samples

taken from the healthy side of the oral cavity of patients with tumors. The division was done to get an indication if the networks could find MAC in the samples. The dataset is described in Table 1 and Figure 1.

2.1.2 CerviSCAN

This dataset is a result from the CerviSCAN project at Uppsala University. From 82 graded pap-smears more than 900 images, each with a focus stack of 41 images, were captured. The microscope used to preform the image acquisition was an Olympus BX51 bright field supplied with a 40x, 0.95 NA objective, resulting in a pixel size of 0.25 μm. To be able to capture focus stacks for each image the microscope was equipped with an E-662 Piezo server controller and actuator. This managed to capture the focus stacks with a step length of 0.4μm. After the images were captured they were manually examined by a cytologist with 30 years experience of screening pap-smears. The cytologist examined each image and marked individual cells and diagnosed these according to the Bethesda system [9]. The resulting dataset can be seen in Table 2 and Figure 2

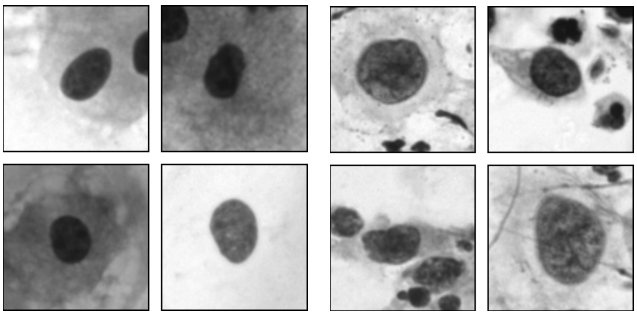
Table 2: CerviSCAN Dataset

Normal

Diagnosis	Nr of cells
Negative for Intraepithelial Lesion or Malignancy	9809

Abnormal

Diagnosis	Nr of cells
Low-grade Squamous Intraepithelial Lesion	766
High-grade Squamous Intraepithelial Lesion	718
Squamous Cell Carcinomas	750



(a) Normal cells from CerviSCAN. (b) Abnormal cells from CerviSCAN

Figure 2: Example images from the CerviSCAN dataset

2.1.3 Herlev Dataset

The Herlev dataset is a publicly available dataset developed in cooperation between the department of Pathology at the Herlev University Hospital and the department of Automation at the Technical University of Denmark [3]. This dataset was created for feature extraction and classification purposes, but not specifically for CNNs [11]. The dataset contains images of varying sizes, with images of normal cells typically including a large cytoplasm while images of abnormal cells typically only show cell nuclei (see Figure 3). As a result, image area as well as the ratio of nucleus size to image size (as shown in Figure 4) may bias classification results. Here, we chose to resize the image using bilinear interpolation, which also impacts the results since interpolation changes the pixel values. Since the abnormal cell images are in general smaller, resizing all images to the same size magnifies these cells more than the normal cells.

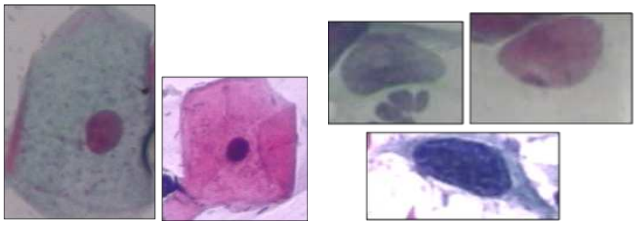
Table 3: Herlev Dataset

Normal

Diagnosis	Nr of cells
Superficial squamous epithelial	74
Intermediate squamous epithelial	70
Columnar epithelial	98

Abnormal

Diagnosis	Nr of cells
Mild squamous non-keratinizing dysplasia	182
Moderate squamous non-keratinizing dysplasia	146
Severe squamous non-keratinizing dysplasia	197
Squamous cell carcinoma in situ intermediate	150

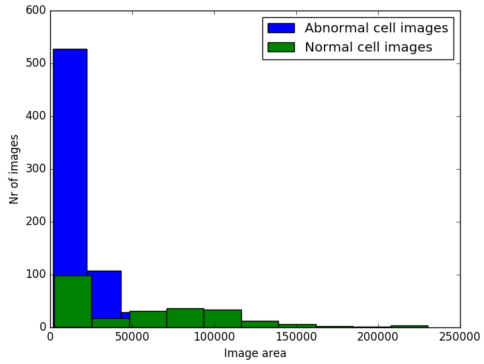


(a) Normal cells from the Herlev dataset. (b) Abnormal cells from the Herlev dataset

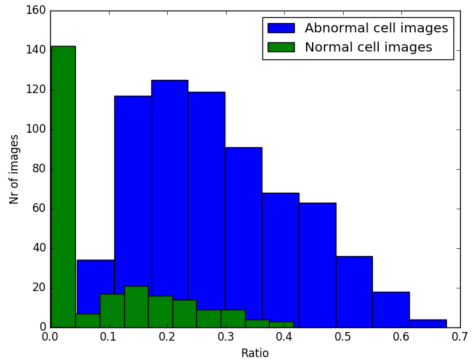
Figure 3: Example images from the Herlev dataset

2.2. Image Preprocessing

Both the Oral and CerviSCAN dataset contains focus stacks for each cell image. To find the image with best focus for each cell the variance of the Laplacian was used. The second order derivative expressed in the Laplacian is known for passing high frequencies which can be an indication of



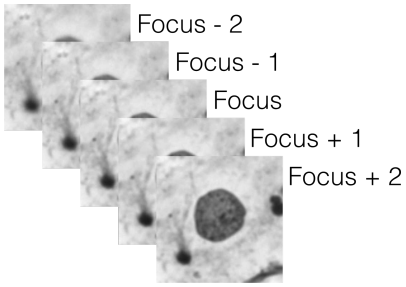
(a) Image sizes.



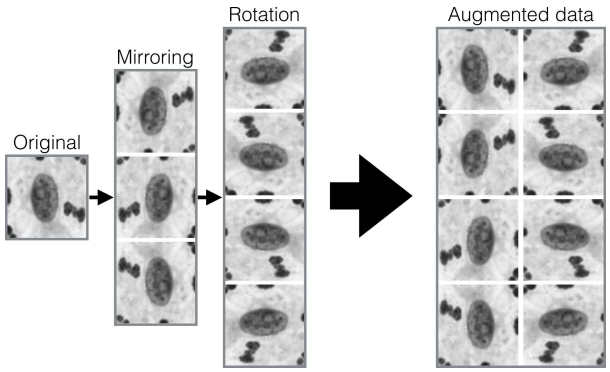
(b) Ratio of nucleus size and image size.

Figure 4: Comparison of different image measurement for the Herlev dataset.

sharp edges in an image [12]. All cells were cut out with a size of 100×100 pixels. The image size was chosen so that the nucleus would fit in the image for all cells. All images of the same cell in the stack were then convolved with the Laplacian and the variance was calculated. To obtain extra information about the cells and not lose information due to bad focus, smaller stacks of each cell were created and used for training and evaluation. These cell images were created using the focus information obtained with the Laplacian. With the image having best focus as the middle image of the stack, the four most adjacent images were stacked below and above to create an image with a depth of five. This means images with a size of $5 \times 100 \times 100$ (Figure 5a). Each image in the image stacks was normalized separately by subtracting the mean and dividing by the standard deviation. To expand the dataset the images were augmented. Augmentation is a regularization technique, that have been proved to improve the results [4]. Since the diagnosis of a cell is depending on the relationship between neighboring pixel values the augmentation was done without any interpolation. The images were mirrored and rotated 90 degrees resulting in eight times as many images per image (Figure 5b).



(a) Example of five depth image used for both the oral dataset and CerviSCAN.



(b) Augmentation.

Figure 5: Input image stack and augmentation methods.

2.3. Model Evaluation

One can look at the network predictions in different aspects, depending on what the aim is. One way is to divide all predictions into four categories: Correctly classified samples (true positive, tp), correctly classified samples that do not belong to the class (true negative, tn), samples that were incorrectly assigned to the class (false positive, fp) and samples that belongs to the class but were not correctly classified (false negative, fn) [16].

With these four classes we calculate how good the network predictions are. Accuracy is the tp + tn to total cell rate, precision represents the tp to tp + fp rate, while recall measures the tp to tp + fn rate. The F-score represents a harmonic mean of recall and precision. The F-score is high only when recall and precision are high. To get a high F-score the network needs to have a low rate of fn and fp [16].

With the oral dataset, to obtain an independent evaluation of the networks the patients were separated for training and evaluation. This means that samples from a given patient were either used for training or evaluation. This ensures that the networks are evaluated on completely unseen data.

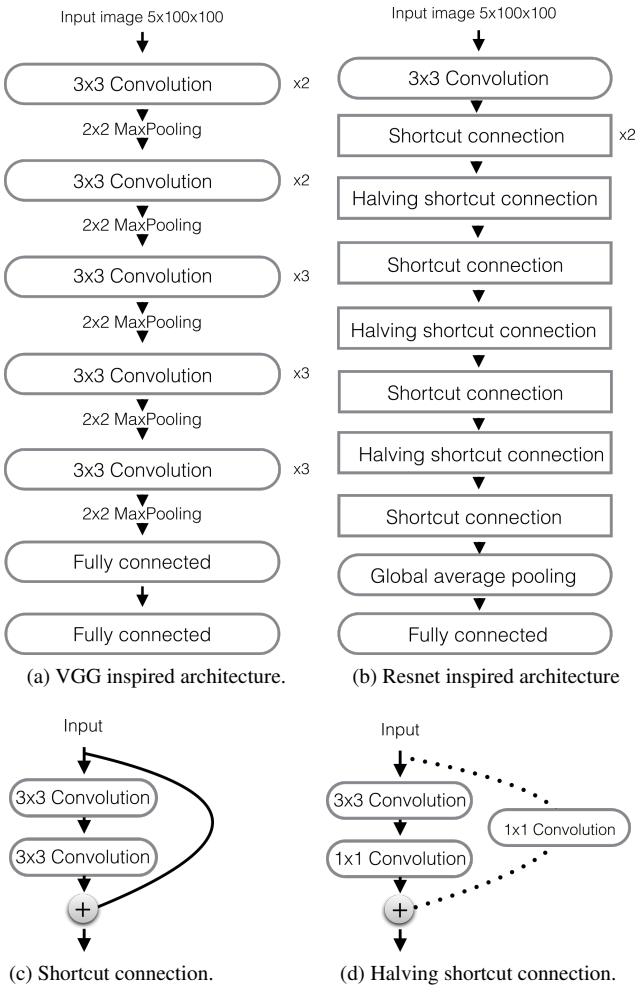
2.4. CNN Architecture and Training

The performance of CNN-based cell classification on the datasets was evaluated on two different network architectures. One was a version of a VGG network [15] and the other a version of ResNet [5]. The VGG architecture used was inspired from the original VGG16 network with 16 weight layers [15]. The main difference is that one fully connected layer (FC) is removed and batch normalization [7] and dropout [17] are inserted to regularize the network. Batch normalization is inserted after every convolutional layer and FC layer. After batch normalization a ReLU layer is inserted as non-linearity. Between the two FC layers dropout layers are inserted with a probability of 0.5, one before batch noramlization and one after. At the end of the network softmax is used to calculate the probabilities.

The ResNet architecture is inspired by the ResNet18 network created by He et al. [5]. In the shortcut connections batch normalization layers are inserted after both convolutional layers. ReLU layers are inserted after the first convolutional layer and after the addition. In the halving shortcut connections batch normalization is inserted after the two convolutional layers. ReLU is inserted after the first convolution and after the addition. Softmax is used at the end of the network to calculate probabilities.

The size of the networks were determined by experimenting with the number of outputs from the first convolutional layer. With each max pooling layer or halving shortcut connection the spatial dimensions are halved while the number of output feature maps are doubled. The optimal value for the number of outputs was found to be 16. This means the first convolutional layer (or shortcut connection) had 16 outputs, the second 32 and so on, ending with 256 outputs for VGG before the FC layers and 128 outputs for ResNet before the global average pooling layer (Figure 6).

The loss in the networks was calculated using cross entropy. The optimization of the networks was done using Adam optimization [8] with a learning rate of 0.01. Each network was trained for 40 epochs and validated once per epoch. The networks were evaluated using K-fold cross validation.



(a) VGG inspired architecture.

(b) Resnet inspired architecture

(c) Shortcut connection.

(d) Halving shortcut connection.

Figure 6: Illustration of the two networks used.

3. Results and Discussion

3.1. Oral Datasets

Since no medical expert has gone through the glasses and selected interesting cells there is no way of knowing that important cells, indicating cancer, have been selected. However, with the assumption that malignancy associated changes might be present in all cells in vicinity of a tumor, the presented result indicates that these changes can be caught by CNNs (Table 4, 5). The patients with tumor were rotated in training and testing to get a K-fold cross validation. The value for K differed for the two datasets. In Oral dataset 1, K = 3 was used and for Oral dataset 2, K = 2 was used. As can be seen in Tables 4 and 5 both networks have a higher score for Oral Dataset 2. An explanation for this can be that the Oral Dataset 1 contains three different patients whilst Oral Dataset 2 only contains two patients. Oral Dataset 1 will then have a larger variation of cells during training, which might lower the performance of the network, but makes the results more stable. The networks were only tested on one single patient per class and more extensive test on multiple patients would be required to draw more general conclusions.

Table 4: 3-Fold evaluation for the Oral Dataset 1

Network	Accuracy	Precision	Recall	F score
VGG	80.66 ± 3.00	75.04 ± 7.68	80.68 ± 3.05	77.68 ± 5.28
ResNet	78.34 ± 2.37	72.48 ± 4.46	79.00 ± 3.37	75.51 ± 3.17

Table 5: 2-Fold evaluation for the Oral Dataset 2

Network	Accuracy	Precision	Recall	F score
VGG	80.83 ± 2.55	82.41 ± 2.55	79.79 ± 3.75	81.07 ± 3.17
ResNet	82.39 ± 2.05	82.45 ± 2.38	82.58 ± 1.92	82.51 ± 2.15

To illustrate a real world implementation of how this method could be used, the networks were tested on individual samples, shown in Figure 7, 8, 9, 10. One can see that healthy patient samples contains few malignant cells and tumor patient samples contains a large amount of malignant cells according to the trained networks. By calculating some form of statistical threshold, a patient could be diagnosed using these results. By comparing the results for VGG and ResNet one can see that VGG classifies more tumor cells as malignant than ResNet, but VGG also classifies more healthy cells as malignant than ResNet. This indicates that the trained VGG networks have a bias towards classifying cells as malignant. If one choses to use ResNet a statistical threshold would then be 30% (Figure 7, 9). Meanwhile VGG would need a statistical threshold at 44%.

3.2. Cervical Datasets

Both datasets have an accuracy and F-score over 84%, which indicates that there is a high potential of detecting cellular changes due to malignancy. There is lower variation in the result for the CerviSCAN dataset compared to the Herlev dataset (Tables 6 and 7). One can also see that the ResNet architecture has a lower standard deviation for both the datasets. With this information it might be possible to showcase a specific amount of cells that, the network predicts, are most malignant which a doctor then could screen. This would decrease the time spent on each sample and thus make it cheaper and more efficient.

Table 6: 5-Fold evaluation for the CerviSCAN dataset

Network	Accuracy	Precision	Recall	F score
VGG	84.20 ± 0.86	84.35 ± 0.97	84.20 ± 0.86	84.28 ± 0.91
ResNet	84.45 ± 0.46	84.64 ± 0.38	84.45 ± 0.47	84.55 ± 0.41

Table 7: 5-Fold evaluation for the Herlev dataset

Network	Accuracy	Precision	Recall	F score
VGG	86.56 ± 3.18	85.94 ± 6.98	79.04 ± 3.81	82.16 ± 3.85
ResNet	86.45 ± 3.81	82.35 ± 5.11	84.45 ± 2.16	83.36 ± 3.65

4. Conclusion and Future Work

Our results on the two cervical datasets are in general terms similar to what has been presented in numerous earlier studies. This simply shows that a CNN with rather limited effort can be trained to reach similar results as can be

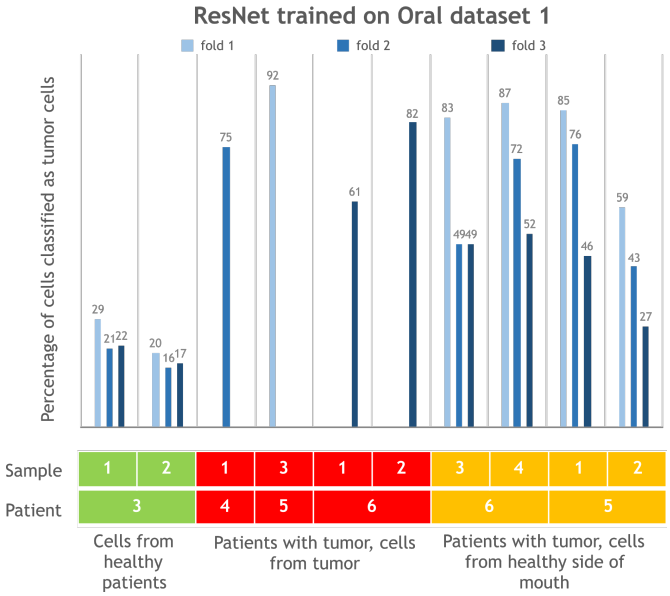


Figure 7: ResNet trained on Oral dataset 1 and evaluated on samples from both datasets.

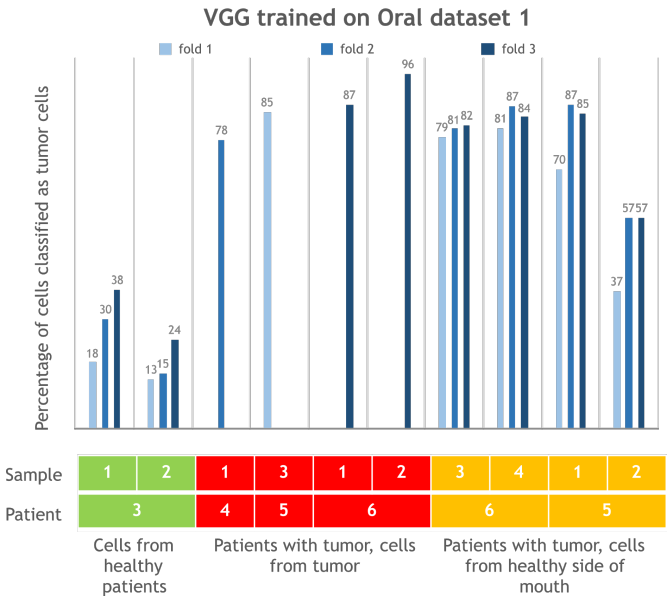


Figure 8: VGG trained on Oral dataset 1 and evaluated on samples from both datasets.

done using much more carefully handcrafted features. The striking result of our small study is that we reached similar performance on the oral dataset in spite of the fact that we there tried to classify all cells, not only selected diagnostic cells, i.e. we applied the MAC approach. This result is even more notable since it was achieved on oral cancer, a cancer type which has been much less studied than cervical cancer.

The results are based on a reasonable number of cells but

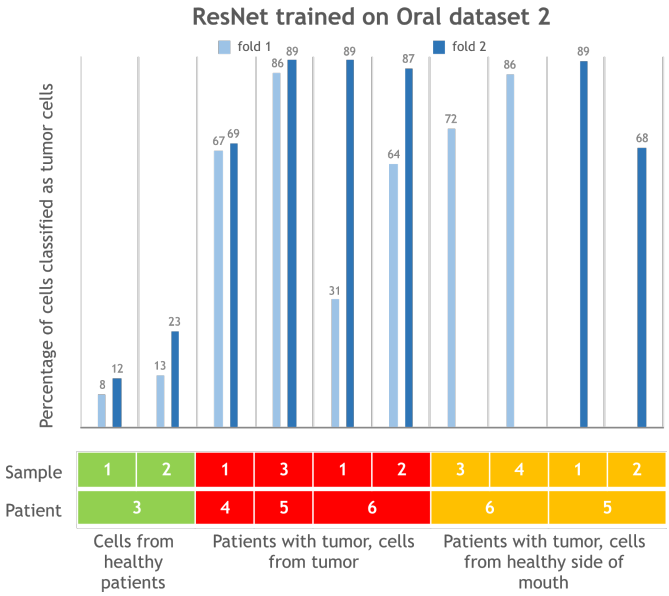


Figure 9: ResNet trained on Oral dataset 2 and evaluated on samples from both datasets.

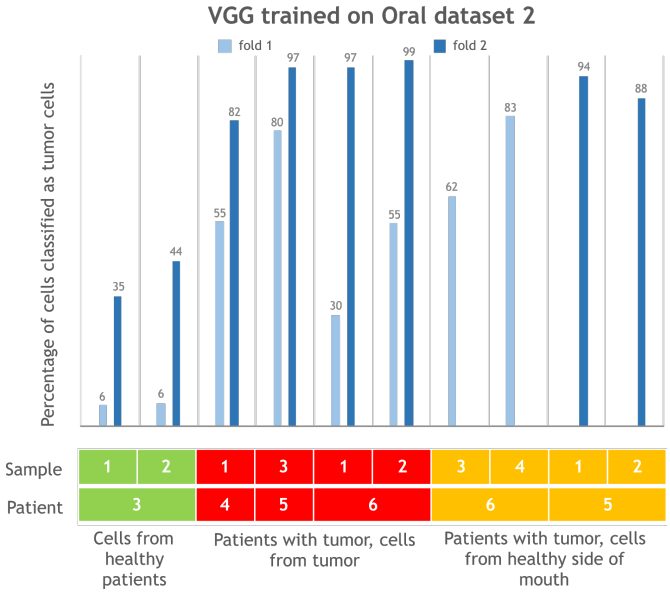


Figure 10: VGG trained on Oral dataset 2 and evaluated on samples from both datasets.

very few patients. The most urgent future work is to extend the material to many more patients. We are currently in the process of planning and organizing such studies.

Between VGG and ResNet one could conclude that ResNet is the preferable network architecture for this classification task. But there are many more network architectures available and new ones are frequently presented so another future possibility is to investigate if some other network architecture could perform the task even better. Since both the oral and cervical cells are similar an interesting future work would be to try transfer learning between the two cancer types. It would also be interesting to see how the results compare to classical image analysis methods.

Acknowledgements

This work was supported by the Swedish research council, grant 2012-4968, ERC Consolidator grant 682810, and Swedish strategic program eSSENCE, to Carolina Wählby. We would also like to acknowledge K. Sujathan of Regional Cancer Center, Thiruvananthapuram, India, for annotating the cervical dataset for this work.

References

[1] E. Bengtsson and P. Malm. Screening for cervical cancer using automated analysis of pap-smears. *Computational and Mathematical Methods in Medicine*, 2014, 2014. doi:10.1155/2014/842037. 1

[2] K. Bora, M. Chowdhury, L. B. Mahanta, M. K. Kundu, and A. K. Das. Pap smear image classification using convolutional neural network. In *Proceedings of the Tenth Indian Conference on Computer Vision, Graphics and Image Processing*, ICVGIP '16, pages 55:1–55:8, New York, NY, USA, 2016. ACM. 2

[3] G. Dounias. Pap-smear (dtu/herlev) databases & related studies. http://mde-lab.aegean.gr/downloads, jul 2008. 3

[4] I. Goodfellow, Y. Bengio, and A. Courville. *Deep learning*. MIT press, 2016. 4

[5] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. 5

[6] J. Hyeon, H. J. Choi, B. D. Lee, and K. N. Lee. Diagnosing cervical cell images using pre-trained convolutional neural network as feature extractor. In *2017 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 390–393, Feb 2017. 2

[7] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *CoRR*, abs/1502.03167, 2015. 5

[8] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. 5

[9] P. Malm. *Image Analysis in Support of Computer-Assisted Cervical Cancer Screening*. PhD thesis, 2013. 2, 3

[10] A. Mehnert, R. Moshavegh, K. Sujathan, P. Malm, and E. Bengtsson. A structural texture approach for characterising malignancy associated changes in pap smears based on mean-shift and the watershed transform. In *2014 22nd International Conference on Pattern Recognition*, pages 1189–1193, Aug 2014. 2

[11] J. Norup. Classification of pap-smear data by transductive neuro-fuzzy methods. Master's thesis, Technical University of Denmark, 2005. 3

[12] J. L. Pech-Pacheco, G. Cristóbal, J. Chamorro-Martínez, and J. Fernández-Valdivia. Diatom autofocusing in brightfield microscopy: a comparative study. In *ICPR*, 2000. 4

[13] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. Berg, and L. Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 12 2015. 2

[14] M. Schwab. *Encyclopedia of cancer*. Springer, 01-01-2012. 2

[15] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. 5

[16] M. Sokolova and G. Lapalme. A systematic analysis of performance measures for classification tasks. *Information Processing and Management*, 45(4):427 – 437, 2009. 4

[17] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958, 2014. 5

[18] WHO. Cervical cancer. http://www.who.int/cancer/prevention/diagnosis-screening/cervical-cancer/en/, 2017. Accessed: 2017-07-03. 2

[19] WHO. Human papillomavirus (hpv) and cervical cancer. http://www.who.int/mediacentre/factsheets/fs380/en/, 2017. Accessed: 2017-07-03. 1

[20] S. W. Wild P.C. *World Cancer Report 2014*. International Agency for Research on Cancer/World Health Organization, 2014. 1

[21] L. Zhang, L. Lu, I. Nogues, R. Summers, S. Liu, and J. Yao. Deeppap: Deep convolutional networks for cervical cell classification. *IEEE Journal of Biomedical and Health Informatics*, PP(99):1–1, 2017. 2