Automatic Classification of Whole Slide Pap Smear Images using CNN with PCA based Feature Interpretation

Kranthi Kiran GV *
kkranthi@student.nitw.ac.in

G Meghana Reddy *
rmeghana@student.nitw.ac.in

National Institute of Technology Warangal, India

Abstract

Classification of whole slide image (WSI) cervical cell clusters traditionally involved two stages including segmentation to crop single cell patches followed by the classification of single cell patches. Hence the performance of classification pipeline depends on segmentation accuracy. We propose a first-time-right method which is a segmentation-free direct classification of WSI cervical cell clusters (without the extraction of single cell patches). The proposed method is evaluated on SIPaKMeD and Herlev datasets. Our method significantly outperformed previous methods and baselines with an accuracy of 96.37% *on WSI patches (cell clusters) and* 99.63% *on single cell images.We also propose a PCA based feature interpretation method to visualize and understand the model to make its decisions more transparent. Our solution is promising in the development of automatic whole slide pap smear image classification system.*

1. Introduction

Cervical cancer accounts for 6.6% of the total positive cases in the world with well over 570,000 cases in 2018 taking it to the fourth spot on the cancer watch list. Approximately 90% of deaths from cervical cancer occurred in low and middle-income countries[15]. The high mortality rate due to cervical cancer can be reduced globally through an approach that includes prevention, effective screening, early diagnosis and treatment programmes. Early and effective screening helps detect precancerous changes which may develop into cancer.

After the introduction of Papanicolaou (Pap) smear[8], the standard screening test for cervical cancer and premalignant lesions is cervical cytology. The analysis of the Pap smear images requires skilled pathologists, and the screening process is expensive and time-consuming. Thus automating this process can help assist the pathologist and provide a less subjective interpretation of the test. There has been considerable work done to automate and create an end-to-end solution for the detection of abnormal cells in the Pap smear slide images which aides the pathologists.

One of the approaches involved moving k-means clustering and SBRG algorithm[4]. This algorithm detects the fine edges of multiple cells in the slide (multiple regions of interest). The algorithm consists of three major steps: (1) threshold values are found automatically using moving k-means clustering algorithm, (2) MSBRG (Modified Seed Based Region Growing) is applied to detect the edges of the region of interest (ROI) based on thresholds, (3) proposed technique is applied to Pap smear images to detect the cytoplasm and nucleus edges of cervical cells.

Apart from edge detection, certain supervised and unsupervised techniques were also explored [10]. Firstly, the locations of nuclei were detected, then a refinement step involving prior information of circumference of the nuclei was performed, and then classification algorithm is applied to detect the abnormal cells in the images. Both supervised techniques such as support vector machines[2] and unsupervised such as fuzzy logic were applied.

All these methods involved preprocessing of the data which includes localisation of the nuclei in the images and application of noise removal algorithms before classification. Researchers have also used Convolutional Neural Networks (CNN) to solve this problem owing to their success in various applications such as image recognition, object detection, segmentation, etc. DeepPap [17] involved (1) detection of the nuclei in the images, (2) cropping of single cell images considering nuclei as the centroid, (3) classifying them using CNNs. This methodology for the detection of abnormal cells in Pap smear images has the following challenges: (1) It depends on the localisation of nuclei, (2) Involved detection and classification of images of single cell whereas classification of the whole slide images is more meaningful and (3) Overlapping cells was a major issue as cropping was done based on the position of nuclei.

* Contributed equally.

Typically, prior to classification, segmentation of the cells in WSI patches into cytoplasm and nuclei is done. But in this paper, we propose a solution by directly classifying the whole slide image patches from pap smear test into normal and abnormal cells. Overlapping cells are no longer a bottleneck to the proposed solution. The inference time is much lesser than the previous work since our method doesn't involve localisation of nuclei and cropping single cells from WSI patches by considering nuclei as centroid prior to classification.

Deep learning models are known to have limited interpretability. In fact, interpretability of deep learning models is a challenging and an active area of research[12, 18]. We propose the use of principal component analysis (PCA) to visualize and understand the features learned by our model to differentiate the classes. We also used saliency maps to make our model predictions more transparent.

To summarise, our contributions are as follows:

1. To the best of our knowledge, we are the first to present a segmentation-free classification of whole slide image patches (without extracting single cell) using convolutional neural networks.
2. We visualize and analyze the feature representations of the model using PCA to provide interpretability and transparency.
3. We show that our approach achieves competitive performance on SIPaKMeD whole slide images. The proposed methodology overcomes issues due to overlapping cells and also achieves significantly improved inference time in comparison to previous methods.

2. Methodology

The proposed methodology consists of three stages: (1) Data preprocessing, (2) Training, (3) Testing.

2.1. Data preprocessing

In our work, we used various data augmentation transformations that are as follows: (1) Since the cervical cells are invariant to rotations, we randomly rotated the input images between -θ to θ with a probability of P_A, (2) We flipped images both horizontally and vertically with a probability of P_B. Unlike most image classification problems, (3) We randomly cropped the image to $H \times W$ size which is also the input size of our network, (4) We applied a random zoom α with a probability of P_A. After applying the above augmentation techniques, we resize the image to $H \times W$.

2.2. Training

We used Residual Net [3] with 34 layers pre-trained on the ImageNet Large Scale Visual Recognition Challenges dataset with 1000 classes achieving a top-1 accuracy of 73.3%. We removed the final classification layer from the network and added a final classification block which consists of a layer that concatenates both average pooling and max-pooling, batch normalization layer, fully connected layer of 512 units (with ReLU activation), batch normalization layer and an output layer with the number of units equal to the number of output classes in the dataset under consideration and retrained it with the corresponding pap smear dataset. This was done to leverage the features learned by the ImageNet pre-trained model. The loss function we used was categorical cross entropy. We used a weight decay of w_d and discriminative learning rates since different layers capture different types of information with the layers closer to input capturing low-level features and the layers closer to output capturing high level features[16]. This allowed us to preserve filters in the layers closer to input that are learned by ResNet on the ImageNet dataset. The network was trained with AdamW Optimizer[7]. The learning rates were scheduled using 1cycle policy which enables Super-convergence allowing faster training of neural networks with very large learning rates and give regularization preventing overfitting[13, 14].

2.3. Testing

During the testing stage, we resized the input image to $H \times W$ size and passed the resized image through the network for it to predict the output class label. For all the tasks on SIPaKMeD dataset and Herlev dataset, we used 5-fold cross-validation with the data splits released along with the dataset.

2.4. Feature interpretation

To understand which features are responsible for the correct prediction of the class of pap smear images, we are analyzing the features of the penultimate layer (ignoring the batch normalization layer and the dropout) of the network i.e the fully connected layer before the output layer. We considered the penultimate layer since it the layer which has the discriminative information which is used by the model to predict. Since they are several features under consideration, we reduced the dimensionality using principal component analysis (PCA). We fit PCA on penultimate layer feature of training data and transform the penultimate layer feature of validation data. We then sorted the validation dataset on the basis of the feature activations obtained by applying the transform on validation set for each feature in the reduced set of features. The analysis of these features can be seen in Figure 2 and Section 4.3.1.

3. Experiments

The following section describes the experimental setup used along with details on datasets used, training strategy, etc.

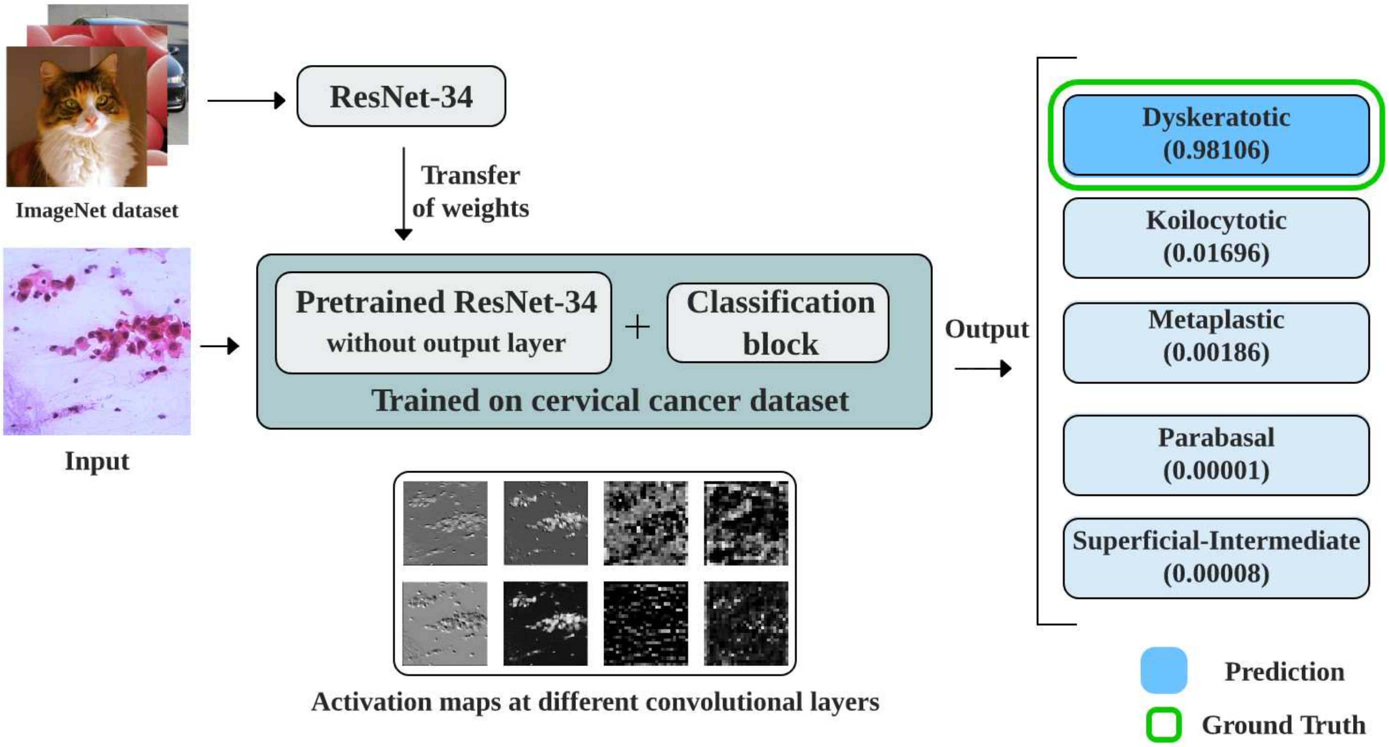


Figure 1. The figure illustrates transfer learning using ImageNet trained ResNet model and the inference pipeline where the model outputs the confidence score of the input belonging to each of the classes. The figure also shows the activation maps of selected convolutional layers for the input image.

3.1. Datasets

3.1.1 Sipakmed

The SIPaKMeD dataset[9] consists of 4049 isolated single cell images which have been manually cropped from 996 cluster cell images of Pap smear slides which we refer to as Whole Slide Image patches (WSI) in this paper. Hence the SIPaKMeD dataset consists of two types of images: (1) whole slide images (2) single cell images. The dataset comprises of cells belonging to five categories: (1) Dyskeratotic, (2) Koilocytotic, (3) Metaplastic, (4) Parabasal and (5) Superficial-Intermediate. 1-2 classes represent the abnormal cervical cells, 4-5 classes represent normal cervical cells, and 3 represents the benign cells.

3.1.2 Herlev

The Herlev dataset[5] consists of 917 isolated single cell images, that is the images contain one cervical cell. There are total of seven classes namely: (1) Superficial squamous epithelia, (2) Intermediate squamous epithelia, (3) Columnar epithelial, (4) Mild squamous non-keratinizing dysplasia, (5) Moderate squamous non-keratinizing dysplasia, (6) Severe squamous non-keratinizing dysplasia and (7) Squamous cell carcinoma in situ intermediate. The classes 1-3 are normal cervical cells whereas classes 4-7 are abnormal cervical cells.

4. Results

4.1. Classification

This subsection describes the experimental setup of classification tasks. All experiments were run on an NVIDIA GTX 1070 GPU.

4.1.1 Hyperparameters and training strategy

In data preprocessing, we used a value of $\theta = 60$, α of 1.0 to 1.1, $P_A = 0.75$ and $P_B = 0.5$. After data processing step, we resized the image to 224×224 $(H \times W)$ in case of WSI patches and 80×80 $(H \times W)$ in case of single cell images. To train the model, we began by freezing the weights of ResNet block of the model and training only the final classification block. The learning rate of the last layer was set 10^{-2}, and each of the previous layers in the final classification block had their learning rate as 10^{-3}. The network was trained for 10 epochs. The weights of the ResNet block were later unfrozen and the training was

Dataset type	Method	Sens (%)	Spec (%)	H-mean (%)	Acc (%)	F-score (%)
WSI	AlexNet	**99.29 ± 0.40**	88.23 ± 1.21	93.43 ± 0.60	88.08 ± 1.45	88.15 ± 1.26
	VGG-16	97.95 ± 0.83	95.65 ± 1.79	96.78 ± 0.87	90.15 ± 0.63	90.00 ± 0.92
	Our model	98.04 ± 0.47	**99.92 ± 0.21**	**98.97.00 ± 0.22**	**96.37 ± 0.53**	**96.38 ± 1.97**
Single cell	Deep (Conv) with SVM [9]	-	-	-	93.35 ± 0.62	-
	Deep (FC) with SVM [9]	-	-	-	94.44 ± 1.21	-
	VGG-16	99.36 ± 0.37	97.33 ± 1.22	98.33 ± 0.56	95.17 ± 1.64	95.23 ± 0.83
	Our model	**99.79 ± 0.19**	**99.83 ± 0.40**	**99.79 ± 0.20**	**99.63 ± 1.1**	**99.63 ± 0.19**

Table 1. Comparison of the performance metrics of our model with previous methods and baselines. The table consists of metric results for WSI patches and single cell images of SIPaKMeD dataset. The highest value for each metric is shown in bold. All experiments followed 5-fold cross-validation.

Method	k-fold CV	Sens(%)	Spec(%)	H-Mean(%)	Acc(%)	F-score(%)
Benchmark [5]	10	98.8 ± 1.3	79.3 ± 6.3	88.0 ± NA	93.6 ± 1.9	-
Bora et al. [1]	5	99.0 ± NA	89.7 ± NA	93.1 ± NA	96.5 ± NA	-
DeepPap [17]	5	98.2 ± 1.2	98.3 ± 0.9	98.3 ± 0.3	98.3 ± 0.7	98.8 ± 0.5
Our model	5	**99.02 ± 0.60**	**99.37 ± 0.21**	**99.15 ± 0.39**	**98.76 ± 0.45**	**99.26 ± 0.42**

Table 2. Comparison of performance metrics, such as Sensitivity(Sens), Specificity(Spec), H-Mean, Accuracy(Acc) and F-score, of our model with previous methods on Herlev dataset. The highest value for each metric is shown in bold.

resumed with lower learning rates between 10^{-5} to 10^{-6} set discriminatively for all the layers. The training of the network was done for a total of 30 epochs. The faster convergence was due to the super-convergence achieved using the 1cycle policy for learning rate scheduling [13].

4.1.2 Evaluation Metrics

Five-fold cross-Validation method was used to report the classification scores for both SIPaKMeD and Herlev datasets. To analyze the performance of the classification model, the following metrics are calculated : (1) Sensitivity (Sens), (2) Specificity (Spec), (3) Accuracy (Acc), (4) H-Mean Score (H-Mean) and (5) F-score. Sensitivity reports the proportion of correctly identified abnormal cells, Specificity reports the proportion of correctly identified normal cells, Accuracy is the overall percentage of correctly identified cells, H-Mean is $2 \times (\frac{Sens \times Spec}{Sens+Spec})$ and F-score is the harmonic mean of precision and recall.

In this section, we have divided the results into two major categories: (1) Quantitative Results, (2) Qualitative Results

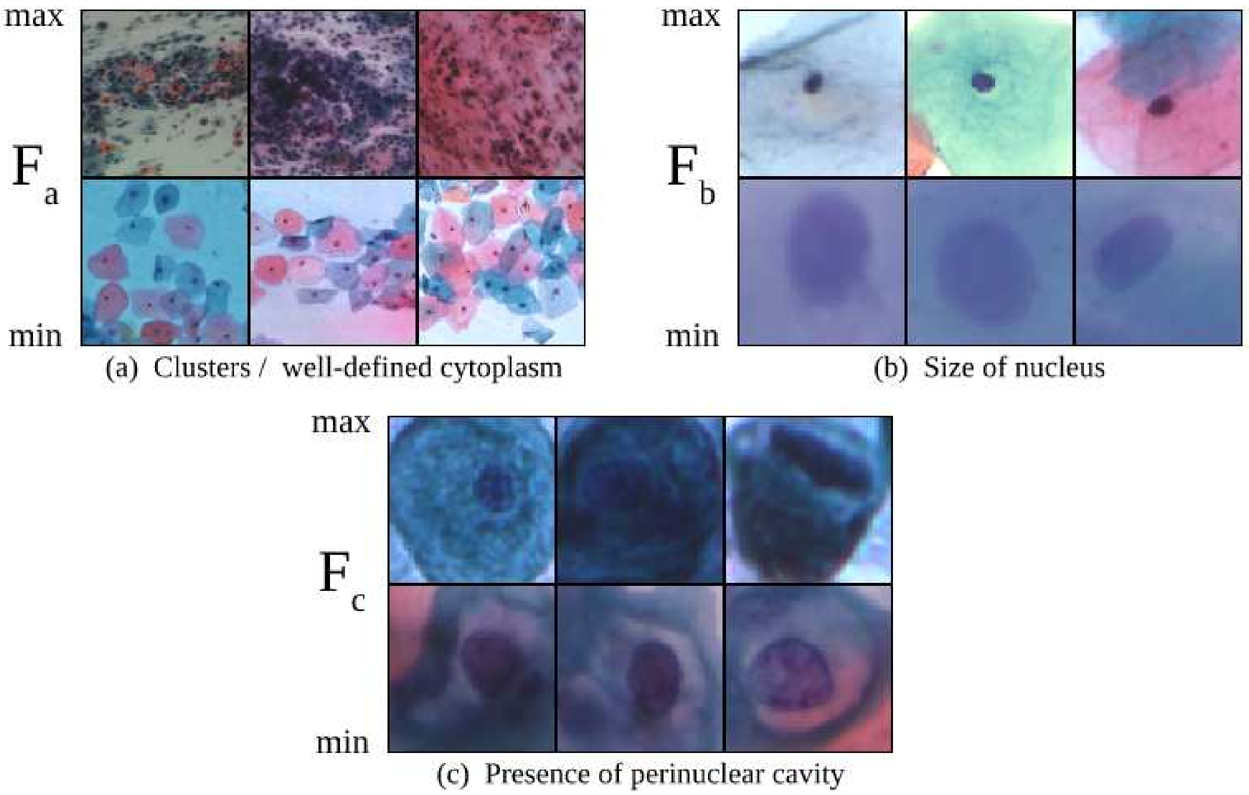
4.2. Quantitative Results

In this section, we describe the classification results on two datasets : (a) SIPaKMeD and (b) Herlev.

Table 1 shows the quantitative comparison of our proposed model with existing methods on WSI patches of SIPaKMeD dataset. Our model performs better in terms of accuracy (96.37%) and F-score (96.38%) in comparison to the AlexNet and VGG-16 baseline models. AlexNet baseline model has a higher sensitivity (99.29%) than ours (98.04%). Although higher *Sens* is desirable even at the expense of lower *Spec*, since WSI patches classified as malignant would be re-examined by human experts, lower *Spec* can increase human labour significantly. Significant portion of WSI patches have normal cells in excess, hence a lower *Spec* results in higher false positives. H-mean balances *Sens* and *Spec* with the model achieving 98.97% in comparison to 93.43 of AlexNet. In experiments conducted on single cell images, the model achieves an accuracy of 99.63% in comparison to 94.44% by Plissiti *et al.* [9].

In Table 2, we observe that our model significantly outperforms the previous methods on single cells images with an accuracy of 99.63%. It also has a higher sensitivity and specificity of 99.79% and 99.83% respectively in comparison to a sensitivity and specificity of 99.36% and 97.33% respectively achieved using VGG-16. A similar trend is observed for other metrics as well.

We can see in Figure 1 that the model predicts the output class correctly for the input image with a high confidence value of 0.98106 for the class Dyskeratotic.



(a) Clusters / well-defined cytoplasm

(b) Size of nucleus

(c) Presence of perinuclear cavity

Figure 2. The figure shows the set of images that have maximum and minimum of a feature activation that is extracted using PCA and then interpreted. The title of each sub-image (a,b,c) is the interpretation of the feature. The sub-image (a) is a WSI patch and (b,c) are single cell cervical images.

4.2.1 Computational Speed

The average training time for 30 epochs is about 15 minutes. This is due to the super-convergence achieved using 1cycle policy[13, 14]. Our proposed method achieves high accuracy without relying on any test time augmentation (TTA). The inference time is 0.0410 sec/image on average which is significantly lower than the existing state-of-the-art system, DeepPap, which has an inference time of 3.5 sec/image (with TTA) and 0.035 sec/image (without TTA but with 1% lower accuracy than with TTA). Our experiments were conducted on NVIDIA GTX 1070 GPU (1920 CUDA Cores) which is less powerful than their NVIDIA GTX TITAN Z (5760 CUDA Cores).

4.3. Qualitative results

4.3.1 PCA feature interpretation

The reduced set of features obtained after applying PCA were analyzed on the basis of value of activation. A few examples are shown in Figure 2. In Figure 2(a), the feature (F_a) is interpreted as clusters/well-defined cytoplasm. This is an important and discriminative feature since Superficial-Intermediate class cells have a clear cytoplasm and nucleus margin whereas Dyskeratotic and Metaplastic cells are usually in clusters with overlapping cytoplasm and nuclei margins. Similarly, (F_b) represents the size of the nucleus which can be used to distinguish Superficial-Intermediate cells (small nucleus), Parabasal cells (large nucleus) and the others intermediate classes. The feature (F_c) represents the presence of perinuclear cavity (cavity between the nucleus and the cytoplasm of the cell). Koilocytotic cells have the presence of perinuclear cavity whereas cells such as Metaplastic cells have a thick cytoplasm instead. The pathologists often use this in recognizing the class of the cell[6].

4.3.2 Sailency Maps

To understand the output of the classifier, visual saliency was explored. Gradient-weighted Class Activation Map (Grad-CAM)[11] is a type of saliency map highlighting the important regions in the image for predicting the output of the given image. In Figure 3(a), we observe that the proposed model focuses on the cell cluster ignoring the background. In Figure 3(b), it can be seen that the nucleus area has the highest activation indicating that the model focuses on the nucleus to output the prediction.

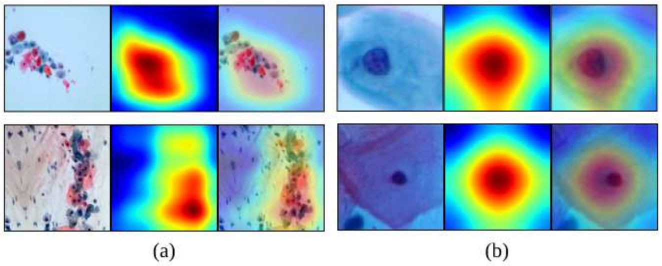


Figure 3. Gradient-weighted Class Activation Maps of (a) WSI patches and (b) Single cell images of SIPaKMeD. In both (a) and (b), the input images are on left, saliency maps are in the middle and the saliency maps overlaid on the input images are shown on right.

5. Conclusion

This paper proposes an end-to-end segmentation-free classification of patches of WSI pap smear images using CNNs. The method is shown to outperform the state-of-the-art methods for cervical cell classification achieving a high accuracy and F-score. We also overcome the limitations of previous works through: (1) No segmentation stage before classification, (2) No bottleneck due to overlapping cells since we classify the WSI patches directly without cropping and (3) Fast inference time of only 0.0410 seconds per image in comparison to 3.5 seconds per image of previous state-of-the-art method [17]. We also showed the transparency of the model by visualizing the features learnt using PCA and saliency maps. By analyzing the penultimate layer of the classification pipeline, we were able to interpret remarkable features learnt by the CNN responsible for the correct prediction.

References

[1] Kangkana Bora, Manish Chowdhury, Lipi B. Mahanta, Malay Kumar Kundu, and Anup Kumar Das. Automated classification of pap smear images to detect cervical dysplasia. *Computer Methods and Programs in Biomedicine*, 138:3147, Jan 2017.

[2] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, Sep 1995.

[3] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, June 2016.

[4] Nor Ashidi Mat Isa. Automated edge detection technique for pap smear images using moving k-means clustering and modified seed based region growing algorithm. *International Journal of The Computer, the Internet and Management*, 13(3):45–59, 2005.

[5] Jan Jantzen, Jonas Norup, Georgios Dounias, and Beth Bjerregaard. Pap-smear benchmark data for pattern classification. In *Proc. NiSIS 2005*, pages 1–9. NiSIS, 2005.

[6] Kitai Kim and Bernard Naylor. *Practical Guide to Surgical Pathology with Cytologic Correlation*. Springer New York, 1992.

[7] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101, 2017.

[8] George N Papanicolaou and Herbert F Traut. The diagnostic value of vaginal smears in carcinoma of the uterus. *American Journal of Obstetrics & Gynecology*, 42(2):193–206, 1941.

[9] Marina E Plissiti, P Dimitrakopoulos, G Sfikas, Christophoros Nikou, O Krikoni, and A Charchanti. Sipakmed: A new dataset for feature and image based classification of normal and pathological cervical cells in pap smear images. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 3144–3148. IEEE, 2018.

[10] Marina E Plissiti, Christophoros Nikou, and Antonia Charchanti. Automated detection of cell nuclei in pap smear images using morphological reconstruction and clustering. *IEEE Transactions on information technology in biomedicine*, 15(2):233–241, 2011.

[11] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, Oct 2017.

[12] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *CoRR*, abs/1312.6034, 2013.

[13] Leslie N. Smith. A disciplined approach to neural network hyper-parameters: Part 1 - learning rate, batch size, momentum, and weight decay. *CoRR*, abs/1803.09820, 2018.

[14] Leslie N. Smith and Nicholay Topin. Super-convergence: Very fast training of residual networks using large learning rates. *CoRR*, abs/1708.07120, 2017.

[15] World Health Organization. Cervical cancer, 2018. [Online; accessed 12-December-2018].

[16] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, pages 3320–3328, Cambridge, MA, USA, 2014. MIT Press.

[17] Ling Zhang, Le Lu, Isabella Nogues, Ronald M Summers, Shaoxiong Liu, and Jianhua Yao. Deeppap: Deep convolutional networks for cervical cell classification. *IEEE journal of biomedical and health informatics*, 21(6):1633–1643, 2017.

[18] Quanshi Zhang, Ying Nian Wu, and Song-Chun Zhu. Interpretable convolutional neural networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8827–8836, 2018.