Cleaning Highly Unbalanced Multisource Image Dataset for Quality Control in Cervical Precancer Screening

Zhiyun Xue[1], Peng Guo[1], Sandeep Angara[1], Anabik Pal[1], Jose Jeronimo[2], Kanan T. Desai[2], Kayode O. Ajenifuja[3], Clement A. Adepiti[3], Silvia D. Sanjose[2], Mark Schiffman[2], and Sameer Antani[1]

[1] National Library of Medicine, Bethesda MD 20894, USA
[2] National Cancer Institute, Rockville MD 20850, USA
[3]Obafemi Awolowo University, Nigeria
zhiyun.xue@nih.gov

Abstract. Automated visual evaluation (AVE) of uterine cervix images is a deep learning algorithm that aims to improve cervical pre-cancer screening in low or medium resource regions (LMRR). Image quality control is an important pre-step in the development and use of AVE. In our work, we use data retrospectively collected from different sources/providers for analysis. In addition to good images, the datasets include low-quality images, green-filter images, and post Lugol's iodine images. The latter two are uncommon in VIA (visual inspection with acetic acid) and should be removed along with low-quality images. In this paper, we apply and compare two state-of-the-art deep learning networks to filter out those two types of cervix images after cervix detection. One of the deep learning networks is DeepSAD, a semi-supervised anomaly detection network, while the other is ResNeSt, an improved variant of the ResNet classification network. Specifically, we study and evaluate the algorithms on a highly unbalanced large dataset consisting of four subsets from different geographic regions acquired with different imaging device types. We also examine the cross-dataset performance of the algorithms. Both networks can achieve high performance (accuracy above 97% and F1 score above 94%) on the test set.

Keywords: Cervical Cancer, Deep Learning, Highly Unbalanced Dataset, Cross-Dataset Evaluation, Anomaly Detection, Acetowhitening, Green-Filter, Lugol's Iodine.

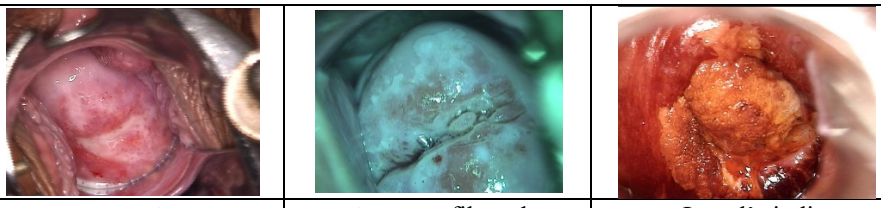
1 Introduction

Cervical cancer affects a significant majority of the developing world where access to clinical care is limited. Having an effective screening program in those regions would reduce the incidence and mortality of cervical cancer significantly. VIA (visual inspection with acetic acid) is a screening method often used in low resource settings. It is an inexpensive alternative to Pap smear and colposcopy test which are methods requiring

significantly higher medical resources with respect to personnel, infrastructure, and devices. In VIA, the cervix before and after the application of a diluted (3-5%) acetic acid is examined by health care practitioners with the naked eye. Based on the visual evaluation results, eligible positive patients can be treated in the same visit with ablation to reduce the loss to follow-up often encountered in low resource regions. While being simple, affordable, and able to provide immediate result and treatment, VIA, which relies on subjective visual evaluation, has moderate sensitivity, specificity, and inter-observer reproducibility [1]. We have recently proposed Automated Visual Evaluation (AVE) using deep learning techniques as an adjunct or a complementary screening method to improve VIA performance [2-4]. AVE has demonstrated promising results on both a large longitudinal population-based dataset of cervigrams (captured using a now-obsolete cerviscope) collected in Guanacaste Costa Rica [2] and a small dataset of images taken by a smartphone enhanced device from several countries [3]. A pilot study has also demonstrated the promise of using deep learning to aid the decision-making process on whether to ablate or not among VIA positives [5]. These results have emboldened a large multi-organizational study initiated and led by the National Cancer Institute (NCI), aiming to improve VIA using automatic computerized algorithms on smartphones or a low-cost specialized handheld device. Due to the adverse impact of COVID-19 on acquiring new data from field studies and the desire of accumulating a large dataset as fast as possible, we have focused on collecting retrospective data from different sources around the world (mainly colposcopy clinics). A side effect of acquiring such retrospective collection is that the data can be of poor visual quality and varied based on the local clinical protocols. Given the huge number of images being processed in the future, automatic handling, in particular, is of high interest. Our studies have aimed to address these using machine learning techniques for image quality control as well as data cleaning in these retrospectively collected datasets. We have developed new algorithms to filter out non-cervix images and blurry images [6,7]. In this paper, we present our work on separating green-filtered and Lugol's iodine cervix images from regular color cervix images taken before or after the application of acetic acid under a white-light illumination.

Cervical cancer is caused by persistent infection from certain high-risk types of HPV (Human Papillomavirus). One main key component in VIA as well as in colposcopy is to observe the color change of the epithelium of the cervix after the weak acetic acid application, a reaction termed as acetowhitening, as HPV infected abnormal tissues may appear more whitish than the neighboring normal squamous epithelium of the cervix. Based on the extent of this reaction (e.g. density, quickness, opacity, thickness), and other visual characteristics, such as vascular patterns, health care providers may assess and grade the severity of the disease. A green filter is commonly used in colposcopy to enhance the visualization of cervical tissue vessels, making suspicious patterns easier to recognize. These green filtered images can be captured by using a green light source, physically attaching a green filter to the camera lens, or created digitally by color manipulation algorithms [8]. Lugol's iodine is another contrast agent applied with the aim of helping identify the lesions overlooked when using acetic acid. Precancerous or cancerous lesions may have different reactions to the iodine solution and may appear in different shades of brown stain from normal tissues, for example, abnormal

tissues could appear to be thick mustard yellow or saffron-colored while normal epithelium tissues would show brown or black color [9]. Fig. 1 shows examples of regular, green-filtered, and Lugol's iodine cervix images respectively.



(a) regular (b) green-filtered (c) Lugol's iodine

Fig. 1. Examples of regular, green-filtered, and Lugol's iodine cervix images.

Although green filter and iodine solution are not usually used in VIA, it is a common practice to use them in colposcopy examinations for visual evaluation of cervix. Therefore, the multi-source datasets we have obtained also contain these two types of images (we call them both non-regular cervix images) and they need to be filtered out. It may appear that these three types of images have distinguished color difference and would be easy to separate, however, there exists a large variance within each type across datasets and there are some images whose color differences across types are subtle and are difficult to differentiate. Further, there is significant variation in the saturation of the green filter which is sometimes also digitally applied as a post-acquisition step. In addition to the goal of automatic cleaning of retrospectively collected data, we want to design the experiments to analyze and study the issues of learning from highly unbalanced dataset and cross dataset/device variance, two main challenges many applications in medical domain face. To the best of our knowledge, this is the first work aiming to investigate these topics for uterine cervix images. The insights gained from this work would help us on applying similar machine learning techniques to other more challenging tasks in AVE.

2 Image Data

We aim to investigate two main issues using these retrospective image datasets: 1) training a model using highly unbalanced data; and 2) examining the model performance across datasets. To this end, in our experiments, we use four datasets which we call the **China dataset**, the **Nigeria dataset**, the **MobileODT dataset** and the **Peru dataset** which were collected from different geographic regions by different providers with different imaging devices. The images within each dataset or across datasets have a large appearance variance with respect to not only cervix or disease related factors (such as woman's age, parity, and cervix anatomy and condition) but also non-cervix or non-disease related factors (such as illumination, focus, specular reflection, presence of clinical instruments, embedded pixel (or graphic) text, imaging device, and variable zoom and angle).

The **China dataset** is a public dataset shared by the authors of [10] on IEEE Dataport. It contains cervix images of 475 patients collected from July 2013 to February 2017 at the First Affiliated Hospital of Science and Technology of China. There are seven images for each subject: one image of pre acetic acid application, four post acetic acid application images taken at different times, one image taken using a green filter, and one image photographed after the application of iodine solution.

The **Nigeria dataset** was collected at the colposcopy clinic of the Obafemi Awolowo University Teaching Hospitals Complex (OAUTHC) in Nigeria by a recent NCI study designed to evaluate the ultimate strategy of AVE triage with self-sampled HPV typing [11]. For each participating woman, multiple images of the cervix were captured using each of the following three image capture devices in order: 1) cellphone, 2) MobileODT EVA, and 3) colposcope mounted with a DSLR camera. MobileODT EVA is a smartphone enhanced device which contains a cross-polarized light source and an external magnifying lens. The images were taken at least one minute after the application of the acetic acid. The data used in this work is the one collected from December 2018 to November 2019. There are no green-filtered images or images with iodine application in this dataset.

In [3], we used a dataset provided by MobileODT to assess whether a deep learning AVE algorithm could perform well on smartphone images. The images were collected from various countries/regions in the world by different providers using MobileODT EVA devices. Each woman had only a single visit, but there were images of varying numbers taken during that visit. The dataset was reviewed for visual quality and images with acceptable/good quality were then annotated by a group of gynecologic oncologists at Rutgers University. The images used in this work are from a subset of the final dataset used to train and evaluate the AVE classifier in [3]. Similar to the Nigeria dataset, this **MobileODT dataset** has no green-filtered images and Lugol's iodine images.

The **Peru dataset** consists of images selected by the collaborating gynecologists in Peru for some of the teaching/training classes on colposcopic visual impression to residents or students. These images were originally collected from the Peruvian Cancer Institute with three different colposcopes. Most of these images have lesions, polyps, or any other important features/characteristics suitable for teaching. Each patient in this dataset has a varied number of images. In addition to acetowhitening images, some patients may have green-filtered images and/or Lugol's iodine images. Some example images in the four datasets are shown in Figure 2.

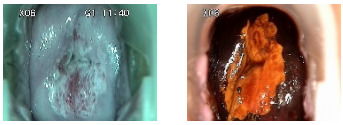
3 Method

As mentioned before, among these four datasets the Nigeria and MobileODT datasets have no iodine and green-filtered images, and only the China and Peru datasets contain such images. To investigate deep networks on high data imbalance and cross-dataset performance, in our experiments, we use China, Nigeria, and MobileODT datasets to train and validate the model and use Peru dataset to test the model. We also set up the problem as separating white light images taken pre or post acetic acid application (named as "regular" images) from green-filtered images and iodine applied images

(named as “non-regular” images), that is, a two-class problem (“regular” vs “non-regular”), since the latter two types of images (“non-regular” images) are usually not used in VIA.

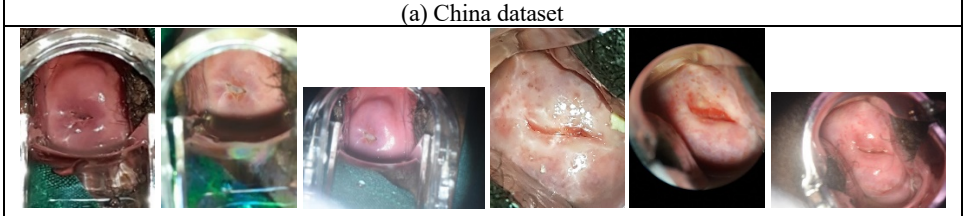
As mentioned previously, there is a large variety in images across datasets and within each dataset. In some images, such as Figure 2(b), there is a significant area outside the cervix region. We have previously developed a cervix detector using RetinaNet [12]. The cervix detector was trained with a dataset other than the images used in this study. Some examples of cervix detection results are shown in Figure 3.

Regular (the first one is pre-acetic acid and the rest are post acetic acid)



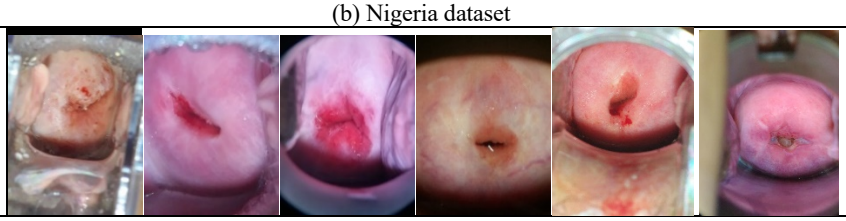
Green-filtered Iodine

(a) China dataset

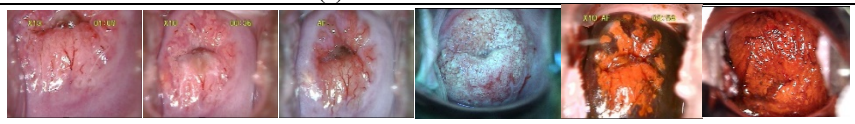


Cellphone EVA Colposcope Cellphone EVA Colposcope

(b) Nigeria dataset



(c)MobileODT dataset



Regular Regular Green-filtered Green-filtered Iodine Iodine

(d) Peru dataset

Fig. 2. Example images in each of four datasets: (a) the sequence of images of one patient in China dataset; (b) images of one patient taken by three devices in Nigeria dataset; (c) images in MobileODT dataset collected from multiple providers in the world; and (d) images of three types in Peru dataset.

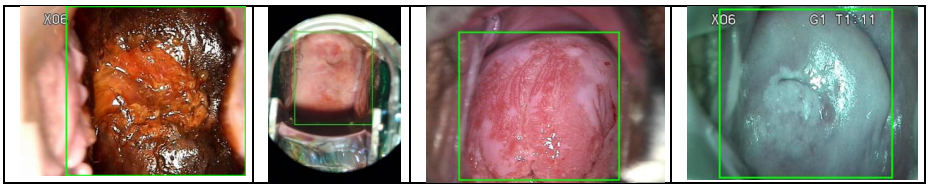


Fig. 3. Cervix detection results shown as a green box overlaid on the images.

Next, we classify the extracted cervix regions into two categories: (a) regular and (b) non-regular images. To this end, we apply and compare two types of deep learning networks. One was designed for anomaly detection (AD), the task of identifying unusual or rare data samples that are significantly different from the majority of data. Our application fits the AD scenario naturally as there are a lot more regular images ("normal") than non-regular images ("abnormal") in our dataset and the non-regular images consists of two different types each of which is significantly different from the large set of regular images. Algorithms of AD [13] use many samples in the normal class to train, either without using any samples from the abnormal class or with using only a small number of abnormal samples. We select and apply Deep SAD [14], a semi-supervised AD deep learning algorithm generalized from Deep SVDD [15] which is an unsupervised one-class classification algorithm. Deep SVDD learns a deep neural network transformation that attempts to map the representations of the unlabeled data (assuming mostly normal) into a "hypersphere" of minimum volume, so that new samples belonging to the normal category would be mapped within this hypersphere whereas the samples belonging to the other category would be mapped outside. Deep SAD is an extension of Deep SVDD which includes a new loss term representing the influence of the additional small amount of labelled data to make the distance of labelled abnormal samples to the hypersphere center further away.

The other deep learning network we apply is a recent ResNet variant network called ResNeSt [16]. ResNet is one of the most popular deep classification networks and is frequently used as a backbone/base network in various networks for computer vision applications. ResNeSt is a variant of ResNet developed with aim of improving the classification performance of ResNet as well as downstream tasks such as object detection and segmentation with comparable computation cost. ResNeSt proposed and included a new block module named "split attention" block. Specifically, in each block, feature maps are first divided into several "cardinal groups", as was done in one previous ResNet variant – ResNeXt [17]. Then, the feature maps in each cardinal group are separated channel-wise into subgroups ("splits"). The features across subgroup splits are combined ("attention") before being concatenated for all the groups. ResNeSt also applies network tweaks and several training strategies (such as augmentation, label smoothing, drop out regularization, and batch normalization) to improve its performance. In our experiments, we test the ResNeSt algorithm with different image input sizes, loss functions and augmentation methods.

4 Experiments and Discussions

As stated previously, four datasets are used for the task of non-regular cervix image filtering. Among them, the China and Peru datasets contain non-regular cervix images, respectively, while the Nigeria and MobileODT datasets contain regular cervix images only. The training set consists of the Nigeria dataset and a part of the China dataset (the original China training set used by the authors for disease abnormality classification in [10]). The validation set consists of the MobileODT dataset and the remainder of the China dataset. The test set consists of the Peru dataset. Table 1 lists the number of patients in each dataset in the training, validation, and test set, respectively. Table 2-Table 4 list the number of corresponding images in both classes in the training, validation, and test sets, respectively. As shown in Table 2, the ratio between the number of regular and non-regular images in the training set is 20746:760 ≈ 27.3, which is relatively high. The images across datasets have varied sizes. All of them are resized to the same size before inputting to the networks.

Table 1. Number of patients in the training/validation/test set

Splits	Dataset	patients
Train	China	427
	Nigeria	988
Validation	China	48
	MobileODT	418
Test	Peru	79

Table 2. Number of training images

Regular		Non-regular	
China (acetowhite)	1895	China (green)	380
Nigeria (cellphone)	8402	China (iodine)	380
Nigeria (EVA)	4815		
Nigeria (colposcope)	5634		
Total: 20746		Total: 760	

Table 3. Number of validation images

Regular		Non-regular	
China (acetowhite)	240	China (green)	48
MobileODT	812	China (iodine)	48
Total: 1052		Total: 96	

Table 4. Number of test images

Regular		Non-regular	
Peru (acetowhite)	307	Peru (green)	58
		Peru (iodine)	43
Total: 307		Total: 101	

For Deep SAD, we adopt the same network architecture used in [14] as the starting point, specifically, the LeNet type CNN for CIFAR-10 in [14] which consists of three modules of 32×(5×5)-filters, 64×(5×5)-filters, and 128×(5×5)-filters, followed by a final dense layer of 128 units. The inverse squared norm loss function is used. For training, the weights are initialized using pretrained autoencoder. The batch size is 4. Adam optimizer with a learning rate of 10^{-4} and weight decay regularization with value 0.5e-6 are used. The number of epochs is set as 150. Figure 4 shows the box plot of the output scores (radius to the hypersphere center) of the images in the validation set, and Figure 5 shows images with high scores and low scores and regular images with high scores after sorting, respectively. Based on the results on the validation set, the threshold for the radius to the hypersphere center is set to be 10. Even though the input image size is small (32 × 32), the model obtains high classification performance on the test

set, achieving sensitivity (recall) 93.1%, specificity 100%, precision 100%, F1 score 96.4%, and accuracy 98.3%, with the cropped cervix images (Non-regular is the positive class). Table 5 lists the corresponding confusion matrix. Figure 6 shows several images misclassified by Deep SAD. They are green-filtered images but are classified as regular images. These images have only a very subtle greenish tone and are hard to distinguish visually unless compared with the regular images from the same patient. Please note, there are some images in the dataset that were zoomed in by the providers to see the region-of-interest when taking the photos and the original images only cover part of the cervix (not due to the failure of cervix detection).

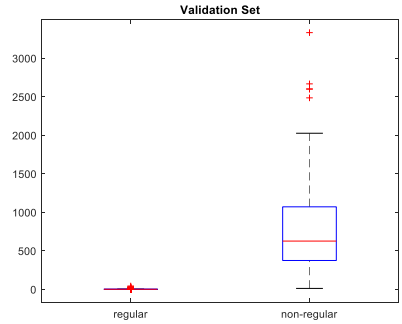
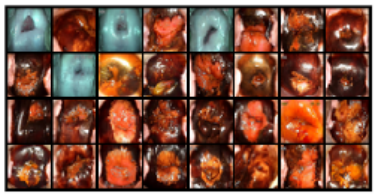


Fig. 4. DeepSAD output scores on val. set

Table 5. Confusion matrix for Deep SAD on the test set

Peru → GT	Non-Regular	Regular
Non-Regular	94	0
Regular	7	307

(a) lowest of all val. images



(b) highest of all val. images

(c) highest of regular images in val. set

Fig. 5. Images in the validation set with DeepSAD low/high scores

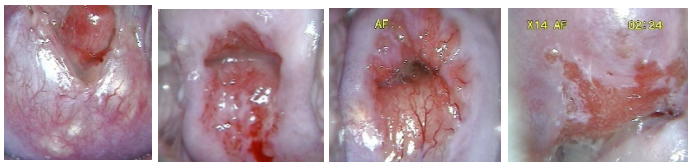


Fig. 6. Misclassified images by DeepSAD

For ResNeSt (specifically, ResNeSt50), we first use the same image size as the one for Deep SAD: 32 × 32. The model is initialized with weights from ImageNet pretrained model. The augmentation methods are random resized crop, random rotation, random horizontal flip and center crop. The loss function is binary cross entropy weighted by the class weight (the total sample size divided by (two times the number of regular/non-regular images)). The optimizer is Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$) with a learning rate of $5x10^{-4}$. The batch size is 256. The model is trained for 100 epochs and the one with the highest performance on validation set is selected. We also train a model with input images of larger size: 224 × 224. For this model, all the above parameters remain the same except the batch size is 64. The performance metrics of all the models on the test set are listed in Table 6. For image size 32 × 32, the performance of ResNeSt model is significantly lower than that of Deep SAD. After increasing the input image size to 224 × 224, the performance (e.g., sensitivity) improves considerably and the F1 score improves to 0.948 from 0.764. The confusion matrix of the ResNeSt models for input size 32 × 32 and 224 × 224 are given in Table 7 and 8, respectively. We also extract the features from the average pooling layer of the ResNeSt models and plot the t-SNE of those features with ground truth labels. As shown in Figure 7 and 8, the ResNeSt (224 × 224) model features in the two categories are separated much better than those from ResNeSt (32 × 32) model.

Table 6. Performance of models

	Sens./recall	Spec.	Prec.	F1	Acc.
ResNeSt, 32 × 32	0.624	0.997	0.984	0.764	0.904
ResNeSt, 224 × 224	0.901	1.000	1.000	0.948	0.976
Deep SAD, 32 × 32	0.931	1.000	1.000	0.964	0.981

Table 7. Confusion matrix for ResNeSt (32 × 32) on the test set

Peru → GT	Non-regular	Regular
Non-Regular	63	1
Regular	38	306

Table 8. Confusion matrix for ResNeSt (224 × 224) on the test set

Peru → GT	Non-regular	Regular
Non-Regular	91	0
Regular	10	307

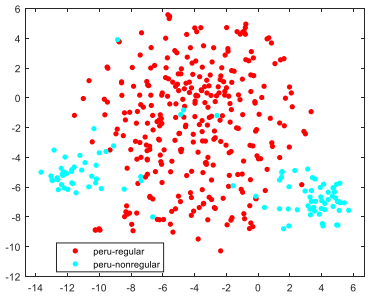


Fig. 7. The t-SNE plot of ResNeSt (32 x 32) features (GT labels)

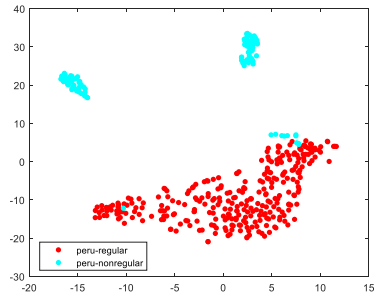


Fig. 8. The t-SNE plot of ResNeSt (224 x 224) features (GT labels)

5 Conclusion and Future Work

In this paper, we present an approach to filter out two non-regular types of images seldomly used in VIA screening of cervical cancer: green-filtered images and iodine-applied images. We are interested in studying data imbalance issue and cross-dataset generalization capacity besides the goal of cleaning retrospectively collected data. To achieve this, we use a combination of several datasets that contain images obtained from four different sources. In the combined dataset, there is large variation with respect to imaging device, patient demography, image quality, illumination source, and existence of clinical obstruction. In addition, there is high imbalance between the number of regular and non-regular images in the combined dataset. To evaluate the model's performance and robustness across multiple datasets, we use images from different sources to train and test in the experiments. We first use an object detection network to identify cervix region. For classification, we apply and compare two types of state-of-the-art deep learning networks: 1) Deep SAD, a semi-supervised anomaly detection network; and 2) ResNeSt, a variant of ResNet. Options such as input image size and loss weighting are compared. High performance can be achieved by both networks (F1 score above 94%). Future work includes testing the model on more datasets, developing a customized Deep SAD network for larger image input size, and removal of other unusable images such as post-treatment cryotherapy images and poor-quality regular images.

Acknowledgement

This research was supported by the Intramural Research Programs of the National Library of Medicine (NLM) and the National Cancer Institute (NCI), both part of the National Institutes of Health.

References

1. Jeronimo, J.; Schiffman, M. Colposcopy at a crossroads. Am. J. Obstet. Gynecol. 2006, 195, 349–353.
2. Hu, L.; Bell, D.; Antani, S.; Xue, Z.; Yu, K.; Horning, M.P.; Gachuhi, N.;Wilson, B.; Jaiswal, M.S.; Befano, B.; et al. An observational study of deep learning and automated evaluation of cervical images for cancer screening. J. Natl. Cancer Inst. 2019, 111, 923–932.
3. Xue, Z.; Novetsky, A.P.; Einstein, M.H.; Marcus, J.Z.; Befano, B.; Guo, P.; Demarco, M.; Wentzensen, N.; Long, L.R.; Schiffman, M.; et al. A demonstration of automated visual evaluation of cervical images taken with a smartphone camera. Int. J. Cancer 2020, 147, 2416–2423.
4. Pal, A.; Xue, Z.; Befano, B., Rodriguez, A.C.; Long, L.R.; Schiffman, M; Antani, S. Deep metric learning for cervical image classification. IEEE Access, vol. 9, pp. 53266-53275, 2021, doi: 10.1109/ACCESS.2021.3069346.

5. Guo, P.; Xue, Z.; Jeronimo, J.; Gage, J. C.; Desai, K. T.; Befano, B.; García, F.; Long, L. R.; Schiffman, M.; Antani, S. (2021). Network visualization and pyramidal feature comparison for ablative treatability classification using digitized cervix images. Journal of Clinical Medicine, 10(5), 953. https://doi.org/10.3390/jcm10050953.
6. Guo, P.; Xue, Z.; Mtema, Z.; Yeates, K.; Ginsburg, O.; Demarco, M.; Long, L. R.; Schiffman, M.; Antani, S. (2020). Ensemble deep learning for cervix image selection toward improving reliability in automated cervical precancer screening. Diagnostics (Basel, Switzerland), 10(7), 451. https://doi.org/10.3390/diagnostics10070451.
7. Guo, P.; Xue, Z.; Long, L.R.; Antani, S. Deep learning for assessing image focus for automated cervical cancer screening. In Proceedings of the IEEE International Conference on Biomedical and Health Informatics, Chicago, IL, USA, 19–22 May 2019.
8. Digiovanni, S. L.; Guaragnella, C.; Rizzi, M.; and Falagario, M. Healthcare system: A digital green filter for smart health early cervical cancer diagnosis. IEEE 2nd International Forum on Research and Technologies for Society and Industry Leveraging a better tomorrow (RTSI), Bologna, Italy, 2016, pp. 1-6, doi: 10.1109/RTSI.2016.7740564.
9. An introduction to colposcopy: indications for colposcopy, instrumentation, principles and documentation of results. Colposcopy and treatment of cervical intraepithelial neoplasia: a beginners' manual, Edited by J.W. Sellors and R.Sankaranarayanan. https://screening.iarc.fr/colpochap.php?lang=1&chap=4.
10. Yue, Z.; Ding, S.; Zhao, W.; Wang, H.; Ma, J.; Zhang, Y.; Zhang, Y. (2020). Automatic CIN grades prediction of sequential cervigram image using LSTM with multistate CNN features. IEEE journal of biomedical and health informatics, 24(3), 844–854. https://doi.org/10.1109/JBHI.2019.2922682.
11. Desai, K. T.; Ajenifuja, K. O.; Banjo, A.; Adepiti, C. A.; Novetsky, A.; Sebag, C.; Einstein, M. H.; et.al. (2020). Design and feasibility of a novel program of cervical screening in Nigeria: self-sampled HPV testing paired with visual triage. Infectious agents and cancer, 15, 60. https://doi.org/10.1186/s13027-020-00324-5.
12. Lin, T.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017, pp. 2999-3007, doi: 10.1109/ICCV.2017.324.
13. Chalapathy, R.; Chawla, S. Deep learning for anomaly detection: a survey. https://arxiv.org/abs/1901.03407.
14. Ruff, L.; Vandermeulen, R. A.; Görnitz, N.; Binder, A.; Müller, E.; Müller, K.; Kloft, M. Deep semi-supervised anomaly detection. The International Conference on Learning Representations (ICLR), 2020.
15. Ruff, L.; Vandermeulen, R. A.; Görnitz, N.; Deecke, L.; Siddiqui, S.A.; Binder, A.; Müller, E.; Kloft, M. Deep one-class classification. Proceedings of the 35th International Conference on Machine Learning, PMLR 80:4393-4402, 2018.
16. Zhang, H.; Wu, C.; Zhang, Z.; Zhu, Y.; Zhang, Z.; Lin, H.; Sun, Y.; He, T.; Muller, J.; and Manmatha, R.; Li, M.; Smola, A. ResNeSt: split-attention networks. https://arxiv.org/abs/2004.08955.
17. Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; He K. Aggregated residual transformations for deep neural networks. https://arxiv.org/abs/1611.05431.