HHS Public Access

Author manuscript

Med Image Learn Ltd Noisy Data (2022). Author manuscript; available in PMC 2022 October 28.

Published in final edited form as:

Med Image Learn Ltd Noisy Data (2022). 2022 September ; 13559: 206–217.

doi:10.1007/978-3-031-16760-7_20.

Image Quality Classification for Automated Visual Evaluation of Cervical Precancer

Zhiyun Xue[1], **Sandeep Angara**[1], **Peng Guo**[1], **Sivaramakrishnan Rajaraman**[1], **Jose Jeronimo**[2], **Ana Cecilia Rodriguez**[2], **Karla Alfaro**[3], **Kittipat Charoenkwan**[4], **Chemtai Mungo**[5], **Joel Fokom Domgue**[6,7,8], **Nicolas Wentzensen**[2], **Kanan T. Desai**[2], **Kayode Olusegun Ajenifuja**[9], **Elisabeth Wikström**[10], **Brian Befano**[11], **Silvia de Sanjosé**[2], **Mark Schiffman**[2], **Sameer Antani**[1]

[1]National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

[2]National Cancer Institute, National Institutes of Health, Rockville, MD 20850, USA

[3]Basic Health International, El Salvador

[4]Department of Obstetrics and Gynecology, Chiang Mai University, Chiang Mai, Thailand 50200

[5]Department of Obstetrics and Gynecology, University of North Carolina-Chapel Hill School of Medicine, Chapel Hill, NC, USA

[6]Cameroon Baptist Convention Health Services, Bamenda, North West Region, Cameroon

[7]Department of Obstetrics and Gynecology, Faculty of Medicine and Biomedical Sciences, University of Yaoundé, Yaoundé, Cameroon

[8]Department of Epidemiology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA

[9]Obafemi Awolowo University Teaching Hospital Complex, Ile Ife, Nigeria

[10]Department of Women's and Children's Health, Karolinska Institutet, Stockholm, Sweden

[11]Information Management Services, Calverton, MD, USA

Abstract

Image quality control is a critical element in the process of data collection and cleaning. Both manual and automated analyses alike are adversely impacted by bad quality data. There are several factors that can degrade image quality and, correspondingly, there are many approaches to mitigate their negative impact. In this paper, we address image quality control toward our goal of improving the performance of automated visual evaluation (AVE) for cervical precancer screening. Specifically, we report efforts made toward classifying images into four quality categories ("unusable", "unsatisfactory", "limited", and "evaluable") and improving the quality classification performance by automatically identifying mislabeled and overly ambiguous images. The proposed new deep learning ensemble framework is an integration of several networks that consists of three main components: cervix detection, mislabel identification, and quality classification. We

zhiyun.xue@nih.gov .

evaluated our method using a large dataset that comprises 87,420 images obtained from 14,183 patients through several cervical cancer studies conducted by different providers using different imaging devices in different geographic regions worldwide. The proposed ensemble approach achieved higher performance than the baseline approaches.

Keywords

Image Quality; Uterine Cervix Image; Automated Visual Evaluation; Mislabel Identification; Ensemble Learning

1 Introduction

Cervical cancer is mainly caused by persistent infection of carcinogenic human papillomavirus (HPV). It is one of the most common cancers among women. Its morbidity and mortality rates are especially high in low- and middle- income countries (LMIC). Besides HPV vaccination, effective approaches to screening and treatment of precancerous lesions play an important role in the prevention of cervical cancer. Precancer is a term that refers to the direct precursors to invasive cancer, which are the main target of cervical screening. In LMIC, due to the limited resources of medical personnel, equipment, and infrastructure, visual inspection of cervix with acetic acid (VIA) is a commonly adopted method for screening for cervical precancer (and treatable cancer). While it is simple, inexpensive, quick to get a result, and does not require expert personnel training, VIA has fairly mediocre intra- and inter- observer agreement and may result in over-treatment and under-treatment [1].

One way to improve VIA screening performance may be to combine it with a low-cost imaging device incorporated with computerized technology that uses predictive machine learning and image processing techniques, called automated visual evaluation (AVE) for the purpose of this discussion [2]. Our proof-of-concept work [2,3] that was demonstrated on two cervical image datasets showed the promise of AVE in LMIC as an adjunctive tool for VIA for screening, or triage of HPV-positive women if such testing is available. Subsequent work has revealed possible problems in implementation [4]. For instance, image quality control, among others, is a key issue.

There are many factors that can adversely affect or degrade image quality. Some are related to clinical or anatomical aspects of cervix, such as the visibility of the transformation zone where cervical cancers tend to arise, and the presence of occlusion due to vaginal tissue, blood, mucus, and medical instruments (e.g., speculum, cotton swab, intrauterine device). Some of these are related to the technical aspect of imaging device and the illumination condition, such as blur, noise, glare, shadow, discoloration, and low contrast, among others. While it is important to train care providers to take high-quality pictures, it is also of importance to develop automated techniques to limit, control, and remedy the image quality problem in existing data sets as well as during acquisition. To this end, we have been working on several aspects, such as filtering out non-cervix images [5], identifying green-filtered images and iodine-applied images [6], separating sharp images from non-sharp images [7], and deblurring blurry images [8]. We also have been working on analyzing

the effects of several image quality degradation factors on the performance of AVE. These include carrying out experiments to quantitively examine and evaluate the AVE results on different levels of image noise and the effectiveness of denoising on AVE [9].

In real-world images, there are often multiple types of degradation existing simultaneously which may vary within as well as across datasets. It may be very difficult to synthesize (mimic) certain types of degradation let alone a combination of multiple degradation types which is significantly harder. Therefore, we have been interested in developing a general image quality classifier using the data labeled by expert clinician annotators based on their judgment. For the quality grading, the annotators were guided by predetermined criteria comprising several factors. They were developed by researchers at National Cancer Institute (NCI) who contributed the first-round quality filtering. Using this guide, NCI researchers assigned image quality labels (“unusable”, “unsatisfactory”, “limited”, and “evaluable”) for images in six datasets that were obtained from different studies, geographical areas, and sources. The guide was specifically designed to reduce the workload and labor time of collaborating gynecologists for annotating images with respect to diagnostic review (for AVE disease grading and treatability analysis), i.e., cutting down the number of low(bad)-quality images among the images to be reviewed by the gynecologists. We noted that the images had a large variance in appearance within each dataset and across datasets. The combined dataset contains 14,183 patients and 87,420 images. We aimed to develop a 4-class quality classifier using this multisource data.

It is common for a large real-world dataset to be noisy and have mis-annotations due to fatigue, misunderstanding, and highly ambiguous samples. Therefore, using all the available training data sometimes may not be the best choice for achieving good generalization. In our dataset, there might be high degrees of ambiguity between some samples in the adjacent classes such as “unsatisfactory” and “limited”, and “limited” and “evaluable” (as implied in the descriptions of labeling criteria in Section 2). We also happened to notice the existence of mislabeled images in the training dataset during random visual browsing. Like labeling itself, manual label cleaning for a large dataset is tedious and labor-intensive. To deal with noisy labels and reduce their negative effects on model performance, one can: 1) design a network that takes weak supervision into consideration; and 2) identify mislabeled or highly ambiguous data automatically. In this paper, we focus on the latter. That is, in addition, to develop an image quality classifier, we are interested in removing/cleaning data used in training to produce better generalization performance.

There are prior works in the literature aiming for mislabel identification. The majority of them monitor the training process and extract certain measures that can be used to represent the difference between clean and mislabeled samples from the training process [10–13]. For example, based on the observation that the curve of the training accuracy with the increase of training epochs is different between clean and bad samples, the authors in [10] developed an iterative approach in which a model was retrained by using only the samples having the lowest loss at the current iteration. Another such example is [11], which proposed a method to use the area under the margin (AUM) value to measure the difference in the training dynamics (as a function of training epochs) between the correctly and incorrectly labeled samples. [11] also developed an effective way (using so-called indicator samples) to find a

suitable threshold value to separate the AUM values of correctly labeled samples from those of in-correctly labeled samples. For our application, we selected and applied a method [14] that is based on an alternative idea. It identifies label errors by directly estimating the joint distribution between noisy observed labels and unknown uncorrupted labels based on the model prediction probability scores [14]. We integrated this algorithm into our image quality classification ensemble framework.

The main contribution of our work is: we developed a new approach that utilizes ensemble methods for both mislabel identification and quality classification for uterine cervix images. We also carried out comparison and ablation experiments to demonstrate the effectiveness of the proposed approach. In the following sections, we first introduce the large dataset collected from multiple sources and the criteria used for manual quality annotation, next present the whole quality classification framework that contains three main components, then describe experimental tests, comparison, and discussion, and at last conclude the paper.

2 Image Quality Labeling Criteria and Data

In this section, we describe the criteria used for AVE image quality annotation and the image datasets that were labeled.

2.1 The Labeling Criteria

These labeling criteria aim to be used for guiding an image taker or health worker for their first round of image quality examination, i.e., to be used for annotating an image based on the technical quality and the ability to see acetowhite areas, not based on anatomical considerations (e.g., squamous-columnar junction (SCJ) observability). There are four image quality categories: *unusable*, *unsatisfactory*, *limited*, and *evaluable*. Images labeled as “limited” or “evaluable” will be used for diagnostic review. The brief guidelines for each category are as follows.

- **Unusable:** The image is one of the following types: non-cervix, iodine, green-filtered, post-surgery, or having an upload artifact.
- **Unsatisfactory**: The image is not “unusable”, but image quality does not allow for evaluation, e.g., has too much blur, is zoomed out/in too much.
- **Limited**: The quality is high enough to allow evaluation of the image, but the image has flaws, e.g., off-center, low light, some blur, obstruction.
- **Evaluable**: The quality is high and there are no major technical flaws. If in doubt, then classify the image as “limited”

2.2 Datasets

The images that were annotated were obtained from 6 cervical cancer studies conducted by different providers with different imaging devices at different regions/countries: NET, Dutch Biopsy (Bx), ITOJU [15], Sweden, Peru, and SUCCEED [16]. The name of the device and the principal investigator (PI) of each study is listed in the Appendix Table 1. For images from the NET study, they were collected from four countries (El Salvador, Kenya, Thailand, and Cameroon) with different image id prefixes. ITOJU study was carried

out in Nigeria and SUCCEED (the Study to Understand Cervical Cancer Early Endpoints and Determinants) was conducted in US. 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, imaging device, clinical instrument, zoom, and angle). In these datasets, one patient may have a varied number of images in one visit or multiple visits. Appendix Table 2 lists the number of annotated images from each study and the number of images in each quality category. In each study, the number of images may vary significantly among quality categories. However, the total numbers of images in the combined dataset across the categories are not highly imbalanced. The reason that Dutch Bx, Sweden and Peru datasets contain a significant percentage of unusable or unsatisfactory images is because many of them are green-filtered or Lugol's iodine applied images or close-up 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. In addition, practitioners in colposcopy tend to take close-up images to check, show or record regions-of-interest, but significantly zoomed-in images are not considered adequate for AVE use as each image is evaluated individually by AVE. A few examples of images in each category are shown in the Appendix Figure 1.

3 Methods

Figure 1 shows the overall diagram of the proposed method. It consists of three main components: 1) cervix detector; 2) quality classifier; and 3) mislabel identifier. The mislabel identifier is based on the result of the quality classifier trained with cross validation. We used three quality classifiers and three corresponding mislabel identifiers. We applied ensemble learning on both the results of mislabel identification and the results of quality classifications. In this study, we aim to remove/clean bad samples from the training and validation sets only, not the test set. The cleaned training and validation sets (the candidates that are identified by all three mislabel identifiers are removed) are then used to train three quality classifiers respectively. The final label of classification is generated by combining the output probability scores of the three classifiers. In the following, we provide more details for each main component.

3.1 Cervix Detection

Since the cervix is the region of interest and the image may contain a significantly large area outside of the cervix, we developed a cervix locator using RetinaNet [17], a one-stage object detection network. We trained the model with a set of images in a different study (Costa Rica Vaccine Trial, conducted by the National Cancer Institute, USA [18]) whose cervices were manually marked and were not used for image quality evaluation/labeling. The detected cervix region was then cropped out and resized before being passed to a classifier. Since not all the images will have a cervix detected (for example, there are images that are not cervix images), all those images in the test set that have no cervix detected by the detector are predicted as "unusable".

3.2 Quality Classification

Image classification has been actively and extensively studied in the literature. Since the debut of AlexNet, many new algorithms or architectures have been developed in this decade of the fast-growing era of deep learning. There are two broad types of neural networks: 1) convolutional neural network (CNN) based, and 2) Transformer based. Using an ensemble of different architectures can utilize the complementary characteristics of the networks and achieve better performance. For our application, we selected three network architectures. Two of those were recent algorithms that have achieved state-of-the-art performance on large general-domain datasets: ResNeSt (ResNet50) [19] and Swin Transformer (Swin-B) [20]. We also added a simpler and smaller ResNet network (ResNet18) for comparison. For all three networks, we initialized their weights using ImageNet pre-trained models and fine-tuned them using our cervix images and labels. To combine the outputs of the three networks, we used the following ensemble method: the class whose average output probability value from all three networks is the largest value among the 4 classes is selected as the final label.

3.3 Mislabel Identification

We applied the algorithm of confident learning (CL) [14] for identifying bad samples in the training data. CL aims to identify mislabels by estimating label uncertainty, i.e., the joint distribution between the noisy and true labels. It uses predicted probability outputs from a classification model for the estimation and is data-centric instead of model-centric. Due to its model-agnostic characteristics, CL can be easily incorporated into our ensemble classification framework. To compute predicted probabilities, K-fold cross-validation is used. In CL, the class imbalance and heterogeneity in predicted probability distributions across classes are addressed by using a per-class threshold when calculating the confident joint [14]. In [21] which uses CL to identify mislabeled images in the ImageNet dataset and all the candidates were reexamined and relabeled by annotators, it showed that many of the candidates were not considered mislabeled by the annotators. Hence, to improve the precision, we used CL for all three classification networks and selected the candidates that were recommended for elimination by all three identifiers.

4 Experimental Results and Discussion

We randomly split the images at the patient level within each dataset into training/validation/test set at the ratio of 70/10/20. Table 1 lists the number of original images in the training/validation/test set in each category in each dataset, respectively. After cervix detection, there were 503, 71, and 141 images that had no cervix detected in training, validation, and test set respectively. Most of these no-cervix-detected images have ground-truth label of “unusable” and a few are of label “unsatisfactory”. For the 141 test images that the cervix detector did not have output, they were all assigned with the prediction label of “unusable” (since the criteria for labeling an image as “unusable” include a “non-cervix” image).

For all the classification models (ResNeSt50, ResNet18, and Swin-B), the input images were resized to 224 × 224, and the weights were initialized with corresponding ImageNet pre-trained model weights. Both ResNeSt50 and ResNet18 models used the same following

hyperparameters: 1) cross-entropy loss with label smoothing; 2) 64 batch size; and 3) Adam optimizer ($\beta1 = 0.9$, $\beta2 = 0.999$) with a learning rate of 5×10–5. For Swin-B, we used: 1) cross-entropy loss function; 2) batch size of 8; 3) Adam optimizer with a learning rate cosine scheduler (initial learning rate was 5×10–5 and the number of warm-up epochs was 5). All three networks used augmentation methods that include random rotation, scaling, center cropping, and horizontal and vertical flip. Each model was trained for 100 epochs and the model at the epoch with the lowest loss value on the validation set was selected.

For identifying bad samples using CL, we created 4-fold cross-validation set using the original training and validation sets and trained 4 models for each classifier. The label uncertainty of both training and validation sets was estimated from their predictions from the 4 models. For pruning, the “prune by noise rate” option was used. The number of mislabeled candidates generated by using each network from the training and validation set is given in Table 2. The number of images in the intersection is much smaller than that generated by any of the networks (6,455 vs. i.e. 13,038).

We used the following metrics to evaluate multi-class classification performance: accuracy and average values of recall, specificity, precision, F1 score, Matthew’s correlation coefficient (MCC), and Kappa score, respectively (using the one-vs-rest approach). Table 3 lists the test set values of the above metrics for each network before and after the removal of identified mislabeled candidates from the training/validation sets. From Table 3, we observed that: 1) when the same set is used to train, the ensemble classifier achieves higher performance than any of the individual classifier; and 2) all the classifiers that are trained with the training data that excludes the mislabeled candidates obtain slightly better performance than those trained using the original data. These observations demonstrate the advantages and effectiveness of ensemble learning, as well as using more data to train may not be helpful and data quality is important. The overall improvement (ensemble plus mislabeled candidate removal) over the best baseline individual model (SwinB) is around 2.7% w.r.t. MCC ((0.683–0.665)/0.665). As shown by [21], some identified candidates may not be indeed mislabeled after the manual re-evaluation. However, to us, it is acceptable to exclude data that are in fact correctly labeled from the training process if it improves the generalization performance.

Figure 2 shows the t-SNE plot of the features extracted from ResNet18 model trained using the original training set as well as the features of the mislabeled candidates and the cleaned training set from the same t-SNE plot. It shows that the cleaned one has a better separation between classes than the original one, indicating the identified candidates may be ambiguous samples. The classification confusion matrix calculated from the test set for the ensemble classifier trained by using the cleaned training/validation set is given in the appendix Figure 2. From the labeling guidelines in Section 2, we expect the main ambiguity to exist between the classes of “evaluable” and “limited” or the classes of “limited” and “unsatisfactory”. It is confirmed by the confusion matrix. As the images predicted with “limited” and “evaluable” will pass the quality check and be used for diagnostic evaluation, we also examined the binary class (“limited+evaluable” vs. “unusable+ unsatisfactory”) classification performance by generating the 2-class confusion matrix from the 4-class one. Its accuracy, F1 score and MCC are: 0.885, 0.859, and 0.762, respectively.

5 Conclusions

The quality of cervix images is important to the succeeding image analysis and visual evaluation for cervix cancer screening. In this paper, we report one of our efforts toward controlling the image quality, i.e., automatically filtering out images of unacceptable quality. To this end, we developed a multi-class classifier using a large, combined dataset that was labeled with four quality categories. Due to factors including ambiguities among classes and the variance in user understanding and interpretation, it is common for a large dataset to have noisy labels. Therefore, we also aimed to improve the generalization performance by identifying and removing bad samples from the training/validation set. By integrating confident learning and ensemble learning, our proposed method achieved better prediction performance than the baseline networks.

Acknowledgement

This research was supported by the Intramural Research Program of the National Library of Medicine (NLM) and the Intramural Research Program of the National Cancer Institute (NCI). Both NLM and NCI are part of the National Institutes of Health (NIH). The NET study was supported partly by Global Good. The authors also want to thank Farideh Almani at NCI for her help.

Appendix

Table 1.

The device and PI of each dataset

	Prefix		Device	PI
NET	El Salvador	Screening population	Samsung J8	Karla Alfaro kalfaro@basichealth.org
	Kenya	Screening population	Samsung J8	Chemtai Mungo chemutai.mungo@gmail.com
	Thailand	Colposcopy clinic	Samsung J8	Kittipat Charoenkwan kittipat.c@cmu.ac.th
	Cameroon	Screening population	Samsung J8	Joel Fokom Domgue fokom.domgue@gmail.com
Dutch Bx	GYFZ	Colposcopy clinic	Digital SLR Camera	Nicolas Wentzensen
ITOJU	HFLD	Screening population	Mobile ODT Eva	Kanan T. Desai and Kayode Olusegun Ajenifuja ajenifujako@yahoo.com kanan_desai2004@yahoo.com
Sweden	PUBG	Colposcopy clinic	Colposcopes	Elisabeth Wikström elisabeth.wikstrom05@gmail.com
Peru	PUBL	Colposcopy clinic	Colposcope	Jose Jeronimo
SUCCEED	SBX	Colposcopy clinic	Digital SLR Camera	Nicolas Wentzensen

Table 2.

Number of images in each dataset in each quality category

	Prefix	Patients	Images	Unusable	Unsatisfactory	Limited	Evaluable
NET	BSPR	82	249	0	24	111	114
	FARH	73	356	0	91	173	92
	JBKV	159	449	3	26	201	219
	ZRQB	157	439	1	6	41	391
Dutch Bx	GYFZ	1036	7886	3839	288	1376	2383
ITOJU	HFLD	1388	19060	177	3991	7633	7259
Sweden	PUBG	878	2221	1072	362	566	221
Peru	PUBL	9736	55082	20423	20820	9568	4271
SUCCEED	SBX	674	1678	14	314	826	524
Total		14183	87420	25529	25922	20495	15474

Fig. 1.
Examples of images in each quality category.

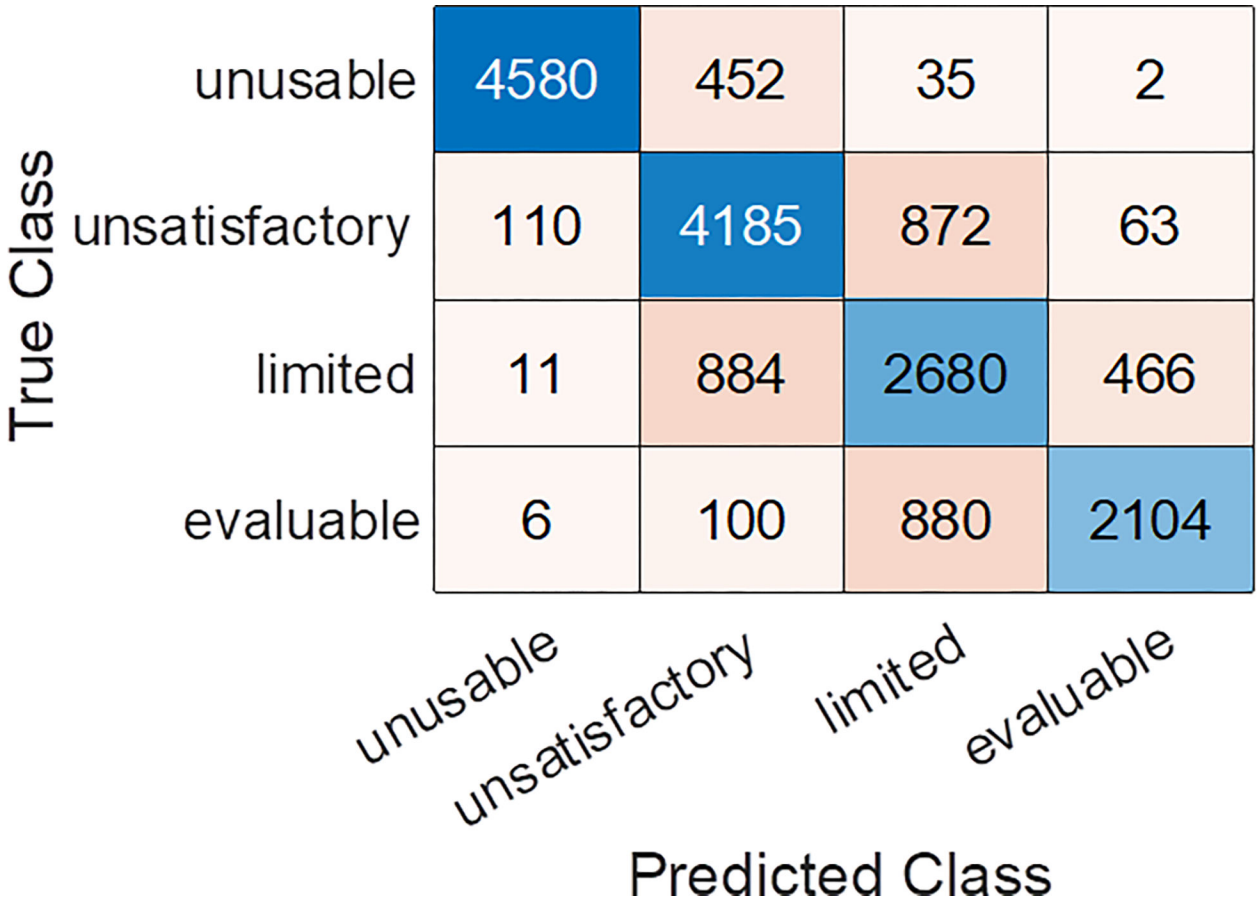


Fig.2.
The classification confusion matrix of the test set

References

1. Jeronimo J, Massad LS, Castle PE, Wacholder S, Schiffman M: Interobserver agreement in the evaluation of digitized cervical images. Obstet. Gynecol, 110, 833–840 (2007) [PubMed: 17906017]
2. Hu L, Bell D, Antani S, Xue Z, Yu K, Horning MP, et al. : An observational study of deep learning and automated evaluation of cervical images for cancer screening. Journal of the National Cancer Institute (JNCI) 111(9), 923–932 (2019). [PubMed: 30629194]
3. Xue Z, Novetsky AP, Einstein MH, et al. : A demonstration of automated visual evaluation of cervical images taken with a smartphone camera. International Journal of Cancer, 10.1002/ijc.33029 (2020).
4. Desai KT, Befano B, Xue Z, Kelly H, Campos NG, Egemen D, Gage JC, Rodriguez AC, Sahasrabuddhe V, Levitz D, Pearlman P, Jeronimo J, Antani S, Schiffman M, de Sanjosé S: The development of “automated visual evaluation” for cervical cancer screening: The promise and challenges in adapting deep-learning for clinical testing: Interdisciplinary principles of automated visual evaluation in cervical screening. Int J Cancer. 2022 Mar 1;150(5):741–752. doi: 10.1002/ijc.33879. Epub 2021 Dec 6. [PubMed: 34800038]
5. Guo P, Xue Z, Mtema Z, Yeates K, Ginsburg O, Demarco M, Long LR, Schiffman M, Antani S: Ensemble deep learning for cervix image selection toward improving reliability in automated cervical precancer screening. Diagnostics (Basel) 10(7), 451 (2020). [PubMed: 32635269]
6. Xue Z, Guo P, Angara S, Pal A, Jeronimo J, Desai KT, Ajenifuja KO, Adepiti CA, Sanjose SD, Schiffman M, and Antani S: Cleaning highly unbalanced multisource image dataset for quality

control in cervical precancer screening. The international conference on Recent Trends in Image Processing and Pattern Recognition (RTIP2R) (2021)

7. Guo P, Singh S, Xue Z, Long LR, Antani S: Deep learning for assessing image focus for automated cervical cancer screening. IEEE EMBS International Conference on Biomedical & Health Informatics (BHI), DOI: 10.1109/BHI.2019.8834495 (2019).
8. Ganesan P, Xue Z, Singh S, Long LR, Ghoraani B, Antani S: Performance evaluation of a generative adversarial network for deblurring mobile-phone cervical images. Proc. IEEE Engineering in Medicine and Biology Conference (EMBC), pp. 4487–4490, Berlin, Germany (2019).
9. Xue Z, Angara S, Levitz D, Antani S: Analysis of digital noise and reduction methods on classifiers used in automated visual evaluation in cervical cancer screening. Proc SPIE Int Soc Opt Eng, 11950:1195008, doi: 10.1117/12.2610235 (2022) [PubMed: 35529321]
10. Shen Y, and Sanghavi S: Learning with bad training data via iterative trimmed loss minimization. In ICML (2019)
11. Pleiss G, Zhang T, Elenberg E, and Weinberger KQ: Identifying mislabeled data using the area under the margin ranking. In Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS). Curran Associates Inc., Red Hook, NY, USA, Article 1430, 17044–17056 (2020).
12. Zhang Z, and Sabuncu MR: Generalized cross entropy loss for training deep neural networks with noisy labels. In Proceedings of the 32th International Conference on Neural Information Processing Systems (NeurIPS) (2018).
13. Patrini G, Rozza A, Krishna Menon A, Nock R, and Qu L: Making deep neural networks robust to label noise: A loss correction approach. In Conference on Computer Vision and Pattern Recognition (CVPR) (2017).
14. Northcutt C, Jiang L, and Chuang I: Confident Learning: Estimating Uncertainty in Dataset Labels. Journal of Artificial Intelligence 70, 1373–1411 (May 2021), 10.1613/jair.1.12125
15. Desai KT, Ajenifuja KO, Banjo A, Adepiti CA, Novetsky A, Sebag C, Einstein MH, Oyinloye T, Litwin TR, Horning M, Olanrewaju FO, Oripelaye MM, Afolabi E, Odujoko OO, Castle PE, Antani S, Wilson B, Hu L, Mehanian C, Demarco M, Gage JC, Xue Z, Long LR, Cheung L, Egemen D, Wentzensen N, Schiffman M: Design and feasibility of a novel program of cervical screening in Nigeria: self-sampled HPV testing paired with visual triage. Infect Agent Cancer 15(60), doi: 10.1186/s13027-020-00324-5 (2020).
16. Wang SS, Zuna RE, Wentzensen N, Dunn ST, Sherman ME, Gold MA, Schiffman M, Wacholder S, Allen RA, Block I, Downing K, Jeronimo J, Carreon JD, Safaeian M, Brown D, Walker JL: Human papillomavirus cofactors by disease progression and human papillomavirus types in the study to understand cervical cancer early endpoints and determinants. Cancer Epidemiol Biomarkers Prev.18(1), pp.113–120. doi: 10.1158/1055-9965.EPI-08-0591 (2009). [PubMed: 19124488]
17. Lin T, Goyal P, Girshick R, He K, and Dollar P: Focal loss for dense object detection. IEEE Transactions on Pattern Analysis & Machine Intelligence 42(02), pp. 318–327 (2020). [PubMed: 30040631]
18. Herrero R, Wacholder S, Rodriguez AC, et al. : Prevention of persistent human papillomavirus infection by an HPV16/18 vaccine: a community-based randomized clinical trial in Guanacaste, Costa Rica. Cancer Discov. 2011
19. Zhang H, Wu C, Zhang Z, Zhu Y, Zhang Z, Lin H, Sun Y, He T, Muller J, Manmatha R, Li M, Smola A: ResNeSt: split-attention networks. https://arxiv.org/abs/2004.08955
20. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S, Guo B: Swin transformer: hierarchical vision transformer using shifted windows," in 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, pp. 9992–10002 (2021).
21. Northcutt C, Athalye A, Mueller J: Pervasive label errors in test sets destabilize machine learning benchmarks. The 35th Conference on Neural Information Processing Systems (NeuIPS) (2021)

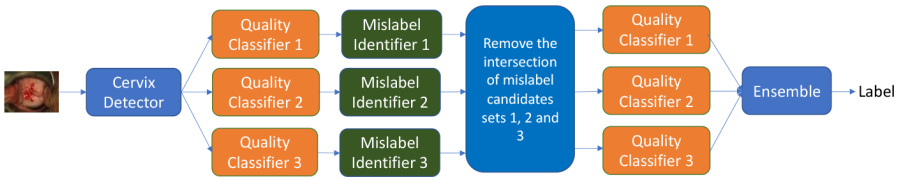
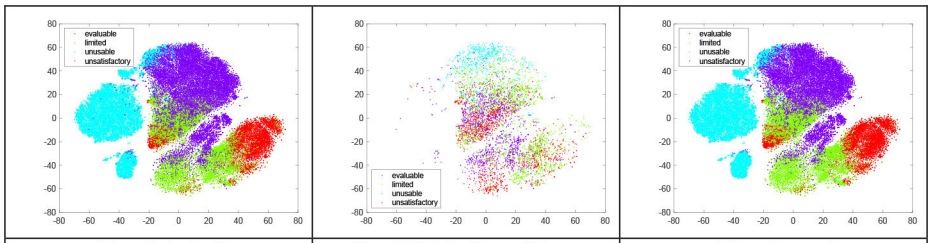


Fig. 1.
Diagram of the proposed approach.



(a) Original (b) Mislabel candidates (c) cleaned

Fig. 2.
T-SNE plots of training set

Table 1.

The number of original images in training/validation/test set.

	Unusable	Unsatisfactory	Limited	Evaluable	Total	No-cervix detected
Train	17734	18083	14376	10836	61029	503
Validation	2726	2609	2078	1548	8961	71
Test	5069	5230	4041	3090	17430	141

Table 2.

The number of identified mislabel candidates.

	Set	Unusable	Unsatisfactory	Limited	Evaluable	Total	
ResNet18	Train	981	3037	4380	3464	11862	13602
	Val.	145	417	664	514	1740	
ResNeSt50	Train	1441	3534	5371	3518	13864	15930
	Val.	197	565	758	546	2066	
Swin-B	Train	1126	3226	4095	2947	11394	13038
	Val.	141	486	625	392	1644	
Intersection	Train	614	1346	1953	1722	5635	6455
	Val.	114	186	257	263	820	

Table 3.

The classification performance on the test set (with/without bad sample removal from the training and validation set).

	Acc	Recall	Spec.	Prec.	F1	MCC	Kappa
Using original train and validation sets							
ResNeSt50	0.742	0.724	0.914	0.734	0.728	0.642	0.650
ResNet18	0.733	0.710	0.910	0.726	0.715	0.628	0.637
Swin-B	0.756	0.744	0.919	0.749	0.746	0.665	0.671
Ensemble	0.766	0.750	0.921	0.760	0.754	0.676	0.682
Using cleaned train and validation sets							
ResNeSt50	0.752	0.734	0.917	0.751	0.740	0.659	0.664
ResNet18	0.746	0.722	0.914	0.747	0.730	0.648	0.654
Swin-B	0.759	0.746	0.920	0.759	0.751	0.671	0.674
Ensemble	0.769	0.752	0.923	0.771	0.759	0.683	0.687