Machine Learning-Based Cervical Cancer Screening Using Cervigrams During Visual Inspection With Acetic Acid: a Systematic Review.

Roser Viñals (roser.vinalsterres@epfl.ch)
EPFL: Ecole Polytechnique Federale de Lausanne https://orcid.org/0000-0002-0043-8235
Magali Cattin
Ecole Polytechnique Federale de Lausanne
Patrick Petignat
Hôpitaux Universitaires de Genève: Hopitaux Universitaires Geneve
Jean-Philippe Thiran
Ecole Polytechnique Federale de Lausanne
Pierre Vassilakos
Ecole Polytechnique Federale de Lausanne

Research

Keywords: cervical cancer, visual inspection with acetic acid, machine-learning, automatic screening

Posted Date: November 16th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-1044067/v1

Abstract

Background: The World Health Organization (WHO) recommendations for promoting effective management of cervical cancer screening in low- and medium-income countries (LMIC) include human papillomavirus (HPV) testing as primary screening followed by visual inspection with acetic acid (VIA) and, if required, treatment. The application of acetic acid induces a transient whitening effect which appears and disappears differently in precancerous lesions and cancer than in benign conditions. However, this assessment by human observers is generally subjective and accuracy is limited. This study presents a systematic review of the automated algorithms for cervical (pre)cancer screening based on images taken during VIA with the objective of assessing their potential as screening tool.

Methods: We performed a systematic literature search in PubMed, Google Scholar and Scopus. The selected studies introduce automated algorithms for the classification of cervical intraepithelial neoplasia grade 2 or higher (CIN2+) with respect to benign conditions, based only on images taken during VIA. We included studies that use, as gold standard, histopathology for CIN2+ cases and, histopathology or normal cytology and colposcopy for benign conditions. The selected studies were analysed in terms of specificity and sensitivity. From each study, the algorithm with the highest accuracy was further studied considering key features such as type of algorithms, acquisition devices, the number of images used per patient, or its performance in comparison to the experts' classification. The quality and risk of the studies was assessed following the QUADAS-2 guidelines.

Results: Of the 1519 studies identified, nine met the inclusion criteria. The algorithms with the highest accuracy from each study reported a sensitivity and specificity values ranging from 0.60 to 0.93 and 0.67 to 0.95, respectively.

Conclusion: Machine learning-based cervical cancer screening algorithms have the potential to become a key tool for cervical cancer screening in countries that suffer from a lack of healthcare infrastructure and personnel. Nevertheless, the selected studies assess their algorithms using small datasets made of highly selected images without reflecting real screened populations. Large-scale and real conditions testing is required to assess the potential of these algorithms as the future of cervical cancer screening.

Systematic review registration: PROSPERO CRD42021270745

Background

In 2020, cervical cancer was the fourth leading cause of cancer death in women. It was the main cause of cancer death in 36 countries, most in sub-Saharan Africa, Melanesia, South America and South-Eastern Asia (1). The World Health Organization (WHO) has defined a global strategy towards the elimination of cervical cancer through intensified vaccination against human papillomavirus (HPV), early detection and effective treatment. This strategy includes screening with high-performance tests of 70% of women aged between 35 and 45 years and treatment of 90% of women identified with precancer or cancer. Projections show that by applying these recommendations by 2030, it may be possible to achieve less than four new cases per 100,000 women worldwide (2).

Less than 30% of low- and middle-income countries (LMICs) have a national HPV vaccination program and only 44% of women in LMICs have ever been screened for cervical cancer. Thus, new screening approaches including sensitive HPV tests are particularly relevant for LMICs. However, the implementation is slow-growing and uneven because of the lack of resources. For this reason, pending the introduction of HPV testing, many countries follow the recommendations of WHO by screening the cervix with visual inspection after application of acetic acid (VIA) which is a simple and inexpensive method. In most instances, women can be tested and eventually treated if positive at the same visit ("see and treat" approach) thereby, overcoming barriers to treatment access (3).

VIA consists of applying acetic acid on the cervix and observing a transient whitening effect appears and disappears differently in precancerous lesions and cancer than in benign conditions. In high-income countries (HICs), the exam is performed with a colposcope, i.e. low-power microscope, used to magnify the view of the cervix and highlight neoplastic abnormalities after the application of acetic acid and Lugol's iodine. This allows experts to rate the cervical appearance as normal or abnormal and guide biopsy sites in women with HPV positive or abnormal cytological results. Nevertheless, colposcopes are rarely available for screening in low-income countries due to limited financial resources, health manpower and facilities. In such conditions, the assessment is usually performed with the naked eye.

Furthermore, it is recognized that visual inspection, with the naked eye or with a colposcope, is often inaccurate for the detection of precancer. Indeed, studies report a sensitivity ranging from 25.0 % to 94.4% for VIA (4–6) and from 39 to 65% for conventional colposcopy (7, 8).

VIA and colposcopy are thus subjective methods associated with intra- and inter-observer variability. Recent advances in artificial intelligence (AI) have been developed and offer great potential for a more objective automated detection of cervical precancer and cancer. AI is already used in healthcare for automated medical diagnoses with smartphones and other devices (9).

In cervical cancer screening, AI-based methods using smartphones, mobile and static colposcopes have been introduced and yielded very promising results (10–18). Still, scant evidence is available on both their usability and impact. Therefore, assessment of the clinical effectiveness

of such tools for automated detection of cervical cancer is necessary, especially in resource-limited settings with a severe shortage of experts.

This study reports the results of a comprehensive systematic review aiming at identifying and comparing AI-based studies that investigate the accuracy of algorithms relying on smartphone and colposcope cervigrams to detect precancer and cancer. The algorithmic methodology of the studies will be also analysed.

Methods

1.1. Protocol and registration

The protocol has been registered in the international prospective register of systematic reviews (PROSPERO CRD42021270745), and reported in compliance with the recommendation of the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA)(19) (Additional file 1).

1.2. Literature Searching strategy

A systematic literature search was done in PubMed, Google Scholar and Scopus. The search strategy included three different concepts: disease (cervical cancer), machine learning-based algorithm (automatic/automate detection, machine learning, deep learning, or artificial intelligence) and acquisition technique (colposcope, colposcopist, colposcopic, visual inspection, or acetic acid). The three concepts were combined using logical operators. The exact searches used in each database and results are shown in Additional file 2. All databases were searched from January 2015 to April 2021.

1.3. Studies selection

Studies satisfying the following criteria were considered eligible. First, the studies should assess an automated algorithm used to distinguish precancerous lesions, named cervical intraepithelial neoplasia grade 2 (CIN2) and grade 3 (CIN3), and cancerous lesions, collectively referred to as CIN2+ (positive), from normal and benign conditions (negative), relying only on images taken during VIA. Studies considering CIN1 as positive or not containing healthy cases were excluded. Second, the gold standard for studying the accuracy of the algorithms should be histopathology results for precancerous and cancerous lesions, which is much more reliable than experts' diagnoses. As conducting biopsies on healthy patients might present ethical concerns, we included studies where negative cases were confirmed by normal cytology and colposcopy or histopathology. However, studies that did not mention their gold standard or for which it was unclear were eliminated. Third, only studies in English from 2015 onwards were included. Finally, only primary peer-reviewed studies were included: review articles, conference abstracts, non-peer-reviewed studies and non-English articles were thus excluded.

1.4. Data extraction and collection

One reviewer (RV) collected the data in a priori-piloted extraction table from PubMed and Google Scholar. For Google Scholar, the software (20) was used to collect the data. Another reviewer (MC) checked the data and performed the same search in Scopus. The assessment of the titles, abstracts and full text-reviews was independently performed by both reviewers (RV and MC). In case of disagreement, a third reviewer was consulted (PV).

From included studies, the study characteristics (e.g. author, publication year, region, database, study design), algorithm description, acquisition device, number of patients included, worst pathological results of each patient, sensitivity, and specificity of the algorithms were extracted directly in a spreadsheet.

1.5. Quality assessment

The Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2) was used to assess the quality of the included studies (21).

1.6. Statistical analysis

The studies included are very heterogenic and each one contains several algorithms based on the same dataset. The performance and methodology of all algorithms presented in the included studies were analysed. For each study, the algorithm with higher accuracy was selected for further analysis.

It is important to mention most of the studies reported their algorithms' performance using cross validation. Furthermore, most of the studies report an average accuracy, sensitivity, and specificity without confidence intervals. Consequently, the analysis and comparison of the studies will be based on the sensitivities and specificities reached by their algorithms, grouping them by year published, acquisition device, number of images used, algorithm family and cross validation technique.

Results

2.1. Study selection

We identified 303, 439 and 989 studies in PubMed, Scopus and Google Scholar, respectively. The papers were filtered by title and abstract, duplicates were removed, resulting in a selection of 84 studies for full-text review. We excluded 75 studies for the following reasons: (i) 19 studies did not mention the gold standard used, (ii) 12 studies classify cervical images in other categories than cervical intraepithelial neoplasia, such as the transformation zone types, (iii) 10 studies used the experts' diagnoses as gold standard, (iv) 9 studies used images during VIA combined with results of the HPV testing, cytological images or visual inspection with other contrast agents such as Lugol's iodine, (v) 8 studies did not use cervigrams taken during VIA, (vi) 6 studies considered CIN1 as positive, (vii) 5 studies did not describe any machine-learning algorithm for screening, (viii) 3 studies were non-peer-reviewed studies and (ix) 3 studies did not contain any normal cases. Finally, 9 studies met all inclusion criteria and thus were included in this review (Figure 1). The included studies are presented in Table 1.

Table 1
Included studies, sorted by publication date.

	Title	First Author	Year	Acquisition device	Cross validation	Acquisition country	Gold standard	Number of patients and images	Number of positive and negative patients
(10)	Adjustable adaboost classifier and pyramid features for image-based cervical cancer diagnosis	T Xu	2015	Film camera (Cerviscope)	10-fold cross validation	Costa Rica	**Negative**: Histopathology **Positive**: Histopathology	**Number of patients**: 1112 patients **Number of images per patient**: 2 sequential images **Number of images used for classifying one patient**: 1 **Training and test sets (sample size)***: 690 patients **Each fold***: 621 patients used for training and 69 for testing	**Imbalanced dataset**: 767 negative (normal or CIN1) and 345 positive (CIN2+). **Prevalence**: 0.31 **Balanced dataset**: 345 negative (normal or CIN1) and 345 positive (CIN2+) **Prevalence**: 0.5
(11)	A new image data set and benchmark for cervical dysplasia classification evaluation	T Xu	2015	Film camera (Cerviscope)	10-fold cross validation	Costa Rica	**Negative**: Histopathology **Positive**: Histopathology	**Number of patients**: 1112 patients **Number of images per patient**: 2 sequential images **Number of images used for classifying one patient**: 1 **Training and test sets (sample size)***: 690 patients **Each fold***: 621 patients used for training and 69 for testing	**Imbalanced dataset**: 767 negative (normal or CIN1) and 345 positive (CIN2+). **Prevalence**: 0.31 **Balanced dataset**: 345 negative (normal or CIN1) and 345 positive (CIN2+) **Prevalence**: 0.5

	Title	First Author	Year	Acquisition device	Cross validation	Acquisition country	Gold standard	Number of patients and images	Number of positive and negative patients
(12)	Multi-feature based benchmark for cervical dysplasia classification evaluation	T Xu	2017	Film camera (Cerviscope)	10-fold cross validation	Costa Rica	**Negative**: Histopathology **Positive**: Histopathology	**Number of patients**: 1112 patients **Number of images per patient**: 2 sequential images **Number of images used for classifying one patient**: 1 **Training and test sets (sample size)***: 690 patients **Each fold***: 621 patients used for training and 69 for testing	**Imbalanced dataset**: 767 negative (normal or CIN1) and 345 positive (CIN2+). **Prevalence**: 0.31 **Balanced dataset**: 345 negative (normal or CIN1) and 345 positive (CIN2+) **Prevalence**: 0.5
(13)	An Observational Study of Deep Learning and Automated Evaluation of Cervical Images for Cancer Screening	L Hu	2019	Film camera (Cerviscope)	Not mentioned	Costa Rica	**Negative**: normal cytology and cervicography **Positive**: Histopathology	**Number of patients**: 9406 patients **Number of images per patient**: Multiple images per patient **Number of images used for classifying one patient**: 1 **Test set (sample size)**: 8917 patients **Each fold**: not applicable	Results reported on 8917 patients: 8689 negative (normal or CIN1) and 228 positive (CIN2+) **Prevalence**: 0.03

	Title	First Author	Year	Acquisition device	Cross validation	Acquisition country	Gold standard	Number of patients and images	Number of positive and negative patients
(14)	A fully-automated deep learning pipeline for cervical cancer classification	Z Alyafeai	2020	Film camera (Cerviscope)	Stratified 10-fold cross validation	Costa Rica	**Negative**: Histopathology **Positive**: Histopathology	**Number of patients**: 348 patients **Number of images per patient**: 1 image per patient **Number of images used for classifying one patient**: 1 **Training and test sets (sample size)**: 348 images **Each fold**: 314 images used for training and 34 for testing	174 negative (normal or CIN1) and 174 positive (CIN2+) **Prevalence**: 0.5
(15)	Classification of cervical neoplasms on colposcopic photography using deep learning	BJ Cho	2020	Colposcope	Not mentioned	South Korea	**Negative**: normal cytology and colposcopy **Positive**: Histopathology	**Number of patients**: 791 patients **Number of images per patient**: 1.8 image per patient in average (1426 in total) **Number of images used for classifying one patient**: 1 **Train set**: 675 images **Test set (sample size)**: 116 images **Each fold**: not applicable	**Training dataset**: 193 negative (normal or CIN1) and 482 positive (CIN2+). **Prevalence**: 0.72 **Test dataset**: 33 negative (normal or CIN1) and 83 positive (CIN2+). **Prevalence**: 0.72

	Title	First Author	Year	Acquisition device	Cross validation	Acquisition country	Gold standard	Number of patients and images	Number of positive and negative patients
(16)	A demonstration of automated visual evaluation of cervical images taken with a smartphone camera	Z Xue	2020	Smartphone	Not mentioned	Various countries in Asia, Africa, North America and South America	**Negative**: Histopathology** **Positive**: Histopathology**	**Data with histopathologic results** **Number of patients**: 537 patients **Number of images per patient**: 2.2 images per patient in average (1159 in total) **Number of images used for classifying one patient**: All images are independently used **Test set (sample size)**: approximately 107 cases (20% of the data are biopsy validated). **Each fold**: not applicable	**Test dataset (biopsy validated)**: 405 negative cases (1027 images) (normal or CIN1) and 132 positive cases (315 images) (CIN2+). **Prevalence (at image level)**: 0.23 **Prevalence (at patient level)**: 0.25
(17)	Diagnosis of cervical precancerous lesions based on multimodal feature changes	G Peng	2021	Colposcope	5-fold cross validation	China	**Negative**: Histopathology **Positive**: Histopathology	**Number of patients**: 300 patients **Number of images per patient**: 2 images (preacetic and postacetic acid) **Number of images used for classifying one patient**: 1 or 2, depending on the algorithm **Training and test sets (sample size)**: 300 patients **Each fold**: 240 patients used for training and 60 for testing	75 normal, 75 CIN1, 75 CIN2, and 75 CIN3 patients **Prevalence**: 0.5

	Title	First Author	Year	Acquisition device	Cross validation	Acquisition country	Gold standard	Number of patients and images	Number of positive and negative patients
(18)	Using Dynamic Features for Automatic Cervical Precancer Detection	R Viñals	2021	Smartphone	Leave-one-out cross validation	Cameroon and Switzerland	**Negative**: Histopathology **Positive**: Histopathology	**Number of patients**: 44 patients **Number of images per patient**: 120 sequential images **Number of images used for classifying one patient**: 120 **Training and test sets (sample size)**: 44 patients **Each fold**: 43 patients used for training and 1 for testing	15 negative (12 normal and 3 CIN1) and 29 positive (11 CIN2 and 18 CIN3) **Prevalence**: 0.66

* Balanced dataset ** Subset of the data with histopathologic diagnoses

CIN: Cervical intraepithelial neoplasia

2.2. Study characteristics

Of the included studies, one collected data in Costa Rica (13), which was also used by different research teams (10–12, 14). One study used data from South Korea (15), one from China (17), one from Cameroon and Switzerland (18) and one did not specify the countries where their data was obtained (16). The images obtained in Costa Rica were taken using a fixed-focus, ring-lit film camera called a cerviscope. During each visit, two sequential images of the cervix after acetic acid application were taken. The studies (15, 17) used colposcopes to acquire the images. In (15), for each participant, several images after the acetic acid application were usually acquired, but only the highest quality image was used for training and testing. In (17), two images were acquired: before and after the acetic acid application. Smartphones were used in (16, 18). In (16), the MobileODT EVA system, which consists of a smartphone and magnifying lens, was used to acquire several images of each patient, after the application of the acetic acid. All images from the same patient were assigned to the same set (train or test) and used independently. In Cameroon and Switzerland, 120 images over 2 minutes were recorded using a smartphone, starting from the moment the acid acetic was applied (18).

The number of patients varies considerably between studies. In (10–12), all studies conducted by the same authors, images from 1112 patients selected from the Costa Rica dataset were used, forming an imbalanced dataset. From the 1112 patients, they randomly selected some cases to create a balanced dataset with 690 cases. In (11), results are reported on both datasets (balanced and imbalanced), while in (10, 12) only the results on the balanced dataset are indicated. For a fair comparison, we will analyse the results of (10–12) only on the balanced datasets. They used 10-fold cross validation: in each fold 621 patients were used for training and 69 for testing. In (13), 9406 patients were included, from which 8917 patients were used for testing their algorithm. The division of the data for training and testing is not clearly detailed. In (14), 348 patients were used, selected from the much larger dataset from Costa Rica. The criteria used for selecting the images are not detailed. Stratified, i.e. the prevalence in the dataset is maintained among folds, 10-fold cross validation was applied: in each fold 314 images were used for training and 34 for testing. In (15), 1426 images from 791 patients were selected after discarding low-quality and blurred images. From each patient, the image with higher quality was selected by two gynaecologic oncologists. Thus, their final dataset contains 791 images. In (16), 7094 images were evaluated by gynecologic oncologists and used for training and testing its algorithm. Additionaly, they tested its algorithm on a dataset with histopathologic results. In total, biopsies were conducted on 537 patients, from which 1159 images were acquired. In (17), 300 cases were used, each one containing a pair of images taken before and after the acetic acid. 5-fold cross validation was used, each fold using 240 patients for training and 60 for testing. Finally, (18) used 120 images from 44 patients and leave-one-out cross-validation at patient level, i.e. 44 folds were done, each one using the 120 images of 43 patients for training and the 120 images of one patient for testing.

Histopathology was used as gold standard in (10–14, 17, 18). In (13, 15), biopsies were only performed when cytology and colposcopic impression after acetic acid was abnormal. Negative cases (normal or CIN1) contained only images from patients with normal cytology and colposcopy. Finally, in (16) only a small subset of the data had histopathological reports. We will only consider this subset.

Most of the studies present several algorithms which are summarized in Table 2. For each study including several algorithms, the algorithm with the highest mean accuracy was chosen for further analysis. Based on the techniques used, we identified three families: traditional machine-learning (ML) techniques, artificial neural networks (ANN), and convolutional neural networks (CNN). Traditional ML techniques are used in the oldest studies: (10, 11), both published in 2015. The rest of the included studies are based on neural networks. Authors in (14, 18) designed small neural networks for the classification task while the rest (12), (13) and (16), (15), (17), adapted popular CNN architectures such as AlexNet (22), Faster R-CNN (23), Inception (24), and VGG16 (25), respectively.

Table 2
Algorithms' description, images used for classification and performance.

	Classifiers	Images used for classification	Mean accuracy	Mean sensitivity	Mean specificity
(10)	AdaBoost classifier. Multi-feature descriptors are used combining a PHOG, PLAB and PLBP*	1 postacetic acid image	0.803	0.864	0.742
(11)	7 models are proposed*				
	(i) RF	1 postacetic acid image	0.800	0.841	0.759
	(ii) GBDT	1 postacetic acid image	0.786	0.820	0.751
	(iii) AdaBoost	1 postacetic acid image	0.768	0.777	0.759
	(iv) SVM	1 postacetic acid image	0.748	0.765	0.730
	(v) LR	1 postacetic acid image	0.742	0.762	0.722
	(vi) MLP	1 postacetic acid image	0.753	0.778	0.728
	(vii) kNN	1 postacetic acid image	0.709	0.751	0.667
(12)	Several classifiers are analyzed. The sensitivity, specificity and accuracy are only specified for few of them*				
	(i) Fine-tuned CaffeNet-based CNN (network adapted from Alexnet (22)) with softmax classifier pretrained with ImageNet	1 postacetic acid image	0.784	0.809	0.759
	(ii) SVM using hand-crafter pyramidal features (PLBP, PLAB and PHOG).	1 postacetic acid image	0.772	0.786	0.758
	(iii) SVM using features extracted with a CaffeNet (network adapted from Alexnet (22)) and one fully connected layer	1 postacetic acid image	0.660	0.651	0.670
	(iv) SVM using features extracted with a CaffeNet (network adapted from Alexnet (22)) and two fully connected layers	1 postacetic acid image	0.691	0.696	0.687
	(v) Fined-tuned SVM using features extracted with a CaffeNet and one fully connected layer	1 postacetic acid image	0.742	0.754	0.730
	(vi) Fined-tuned SVM using features extracted with a CaffeNet and two fully connected layers	1 postacetic acid image	0.746	0.765	0.728
	(vii) Fined-tuned AdaBoost classifier using features extracted with a CaffeNet and two fully connected layers	1 postacetic acid image	0.774	0.809	0.739
(13)	Faster R-CNN architecture (23).	1 postacetic acid image	0.832	0.930	0.830
(14)	4 different models.				
	(i) CNN with two convolutional layers. Automatic feature extraction	1 postacetic acid image	0.682	0.597	0.774
	(ii) CNN with three convolutional layers. Automatic feature extraction	1 postacetic acid image	0.703	0.723	0.683
	(iii) ANN with one hidden layer. Hand-crafted features.	1 postacetic acid image	0.729	0.690	0.768
	(iv) ANN with two hidden layers. Hand-crafted features.	1 postacetic acid image	0.771	0.752	0.780
(15)	2 networks used.				
	(i) Inception-Resnet-v2 (modified version of the Inception-v3 model) (24)	1 postacetic acid image	0.693	0.667	0.706
	(ii) Resnet-152 (updated version of the Resnet model) (26)	1 postacetic acid image	0.689	0.667	0.699
(16)	Faster R-CNN architecture (23)**	1 postacetic acid image	Area under the ROC curve = 0.87 (95% CI 0.81-0.92)		

	Classifiers	Images used for classification	Mean accuracy	Mean sensitivity	Mean specificity
(17)	Algorithms based on VGG16 (25)				
	(i) Network model that uses VGG16 to extract the features of postacetic acid test colposcopy images	1 postacetic acid image	0.660	0.647	0.675
	(ii) Network model that uses VGG16 to extract the features of preacetic acid test and postacetic acid test colposcopy images	1 preacetic acid and 1 postacetic acid image	0.717	0.728	0.707
	(iii) Network model that uses VGG16 to extract the features of registered preacetic acid test and postacetic acid test colposcopy images	1 preacetic acid and 1 postacetic acid image	0.767	0.755	0.774
	(iv) Network model that uses VGG16 to extract the features of registered preacetic acid test and postacetic acid test cervical images (after cervical region extraction).	1 preacetic acid and 1 postacetic acid image	0.863	0.841	0.898
(18)	Pixel-wise classification using an ANN with one hidden layer and combined with region growing segmentation.	120 images during VIA	0.886	0.897	0.867

Selected as representative algorithm of the study (higher accuracy) *Balanced dataset **Subset of the data with histopathologic diagnoses

ANN: artificial neural network; CI: confidence interval; CNN: convolutional neural network; GBDT: gradient boosting decision tree; kNN: k-Nearest Neighbors; LR: logistic regression; MLP: multilayer perceptron; PHOG: pyramid histogram of oriented gradients; PLAB: pyramid color histogram in L∗A∗B space; PLBP: pyramid histogram of local binary pattern; RF: Random forest; characteristic; ROC: receiver operating characteristic; SVM: support vector machine; VIA: visual inspection with acetic acid;

2.3. Sensitivity and specificity analysis

To compare the performance of the algorithms, the sensitivity and specificity have been analysed. All the studies reported the sensitivity and specificity, except (16) which reports the receiver operating characteristic (ROC) curve. From this ROC curve, we consider the accuracy that is maximized with the following formula: $Accuracy = prevalence \bullet sensitivity + (1 - prevalence) \bullet specificity$. Figure 2b shows the sensitivity and false positive rate for all the algorithms, as well as the ROC curve of (16). Note that the algorithms use different sample sizes which is an essential feature to consider. For instance, study (18) achieves high sensitivity and specificity compared to other algorithms but its sample size is only 44 patients, presenting a high risk of overfitting. By contrast, in (13), its algorithm was tested on 8917 patients probably providing more reliable results. To represent the sample sizes, the circle areas are proportional to the sample sizes.

The algorithms with the highest accuracy from all studies are shown in Figure 2b. To compare these algorithms, the plots have been coloured depending on the year published (Figure 2c), acquisition device (Figure 2d), number of images used (Figure 2e), algorithm family (Figure 2f) and cross validation technique (Figure 2g). Finally, a comparison of the algorithms with respect to the experts is shown in Figure 2h.

2.4. Quality assessment

The quality of the included studies has been assessed using QUADAS-2 (Figure 3). We analysed both the risk of bias and the applicability concerns due to flow and timing, reference standard, index test and patient selection.

The risk related to flow and timing, and index test is low for all the included studies. Nevertheless, in eight (10–12, 14–18) out of the nine included studies, the risk related to patient selection is high. Indeed, small datasets are used, with highly selected images and without matching the real rate of the populations screened. The only study with a low risk to introduce bias is (13), whose dataset includes images from more than 8000 patients with a prevalence of 0.03. Most of the included studies used histopathology as gold standard for all the patients. However, studies (13, 16) did only use histopathology as gold standard for positive patients, having a high risk to introduce bias. In (15), only a small subset used histopathology as gold standard. As we have focused our analysis on this subset, we consider that it has low risk. No applicability concerns were identified.

Discussion

The algorithms' performances of the included studies are presented in many different ways, most of them using cross validation and not mentioning confidence intervals nor the number of false positives, false negatives, true positives and true negatives (10–12, 14–17). This lack of confidence intervals prevents objective statistical analysis of the results and the possibility of meta-analysis. The common metrics used in all the included studies except (16) - which provides the ROC curve - are the average accuracy, specificity, and sensitivity. Consequently, we have focused our analysis on the specificities and sensitivities.

The three most recent algorithms are the ones that have a higher specificity (16–18) (Figure 2c). Nevertheless, they only contain 537, 300 and 44 patients, respectively. Using such small datasets could give rise to a lack of generalization of the proposed methods. Further testing in larger datasets should be done to accurately estimate their performance. Similarly, from Figure 2g we can observe that the two algorithms with higher sensitivity also include a limited number of cases (18) or do not mention if they performed cross-validation (13), having also a higher risk of overfitting.

Two out of the three algorithms with the highest specificity use more than one image per patient: in (17), a pair of images taken before and after the acetic acid, and, in (18), 120 sequential images after the application of the acetic acid (Figure 2e), showing the potential of using multiple sequential images of the cervix. Colposcopists not only detect the precancerous lesions based on the intensity of aceto-whitening but also based on its time evolution. Furthermore, (27) demonstrated that although most of the lesions are visible one minute after the application of acetic acid, it is reasonable to perform VIA during 3 minutes. Thus, instead of using single images as done in (10–16) using multiple sequential images could enhance automatic screening algorithms.

Two key features that could enable the use of these tools in LMICs have been identified. Firstly, algorithms that use images from portable devices (smartphone and camera) achieve similar or better performance than the ones using colposcopes (Figure 2e). For LMICs, portable devices seem to be more appropriate for acquiring and analysing the images than expensive tools such as colposcopes. Secondly, simpler algorithms achieve - even outperform – similar performance than some of the CNNs based algorithms (Figure 2f). For instance, in (10–12), all studies conducted by the same authors and using the same database, two traditional ML-based algorithms are presented (10, 11), published in 2015 as well as one CNN-based method, published in 2017, (12). The three studies present similar results even when using much more sophisticated algorithms in (12) than in (10, 11). The use of simpler algorithms facilitates the integration into mid-range smartphones allowing offline use of the tool. By contrast, more sophisticated algorithms such as the ones based on complex architectures might require the use of external servers for performing the classification.

The main limitations of most of the included studies are the limited number of patients or images, the high selection of the patients used to train and test the algorithms, or the lack of large-scale tests. The risk of bias due to patient selection was defined high for (10–12, 14–18) when assessing the risk with QUADAS-2. In Costa Rica, the study (13) recruited a large number of patients resulting in a prevalence of 0.3. The other studies relying on this dataset have chosen a small subset of cases to develop their algorithms, without specifying the criteria for the patients' selection and having a prevalence ranging from 0.31 (10–12) to 0.5 (14).

Out of the 9 studies, only (13) and (15) have a high risk of bias due to the gold standard as they use histopathology only to confirm positive patients. In (13), only patients with abnormal cytology or visual inspection were referred to colposcopy and biopsied(13). Thus, its dataset consists of positive cases confirmed by histology and negative cases confirmed by normal cytology and normal visual inspection. Similarly, in (15), patients with normal cytology and colposcopy were considered negative while positive cases were confirmed by pathology. The remaining studies used histology as gold standard for all the patients or, in (16), for a small subset used for testing.

Furthermore, different screening approaches were used in the included studies to collect the data. During each screening visit in Costa Rica, cytology, HPV testing and visual inspection with acetic acid were performed (10–14). Studies (15), (16) and (17) do not indicate the screening or patient selection criteria. In (18), women were referred for colposcopy after positive HPV testing in Cameroon, and after both positive cytology and HPV testing in Switzerland.

Finally, only two of the included studies, (13) and (18), present a comparison between experts and their algorithms. In (13), each pair of images taken during VIA was graded by one expert, blinded from the histopathologic diagnoses, as normal, atypical, low-grade lesions, or CIN2+. In (18), three experts classified, blinded from the histopathologic diagnoses, the 44 patient's images as positive (CIN2+) or negative. From Figure 2h, we can observe that the algorithms in both studies achieved higher sensitivity than the experts. In (18), on average, experts had worse sensitivity and specificity than the algorithm, while in (13), experts had higher specificity than the automated algorithm in identifying atypical cases or CIN2+ cases.

Conclusion

Automated algorithms for precancerous and cancerous lesions' detection using images acquired during VIA are very promising for countries that suffer from the lack of expensive tools and trained personnel, where most of the deaths caused by cervical cancer occur. These images can be taken by portable devices such as smartphones or cameras, which are more accessible in LMICs than colposcopes, with no observed decrease in the performance of the algorithms.

The performance of AI applications using single or sequential cervical images may be as accurate as – or even better - the human interpretation of the same images. Nevertheless, most of the included studies have assessed their algorithms using small datasets, with highly selected images and without reflecting the real rates in screened populations. Furthermore, most of the studies do not report confidence intervals, preventing a

more objective analysis. Consequently, large-scale, and real condition testing of these algorithms, with proper statistical analysis of their results, is required to assess the potential of these algorithms to become the future of cervical cancer screening.

Abbreviations

AI: artificial Intelligence; ANN: artificial neural network; CI: confidence interval; CIN: cervical intraepithelial neoplasia; CNN: convolutional neural network; GBDT: gradient boosting decision tree; HIC: high-income countries; HPV: human papillomavirus; kNN: k-Nearest Neighbors; LMIC: low- and medium-income countries; LR: logistic regression; ML: machine-learning; MLP: multilayer perceptron; PHOG: pyramid histogram of oriented gradients; PLAB: pyramid color histogram in L∗A∗B space; PLBP: pyramid histogram of local binary pattern; PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses; QUADAS: Quality Assessment of Diagnostic Accuracy Studies; RF: Random forest; ROC: receiver operating characteristic; SVM: support vector machine; VIA: visual inspection with acetic acid; WHO: World Health Organization.

Declarations

Ethics approval and consent to participate: Not applicable

Consent for publication: Not applicable

Availability of data and materials: The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Competing interests: The authors declare that they have no competing interests

Funding: This research has been supported by the Solidarité Internationale Genève and Tech4Dev (École Polytechnique Fédérale de Lausanne).

Authors' contributions: RV, MC and PV designed the study. RV and MC screened the titles, abstracts, and full text-reviews. PV developed the search strategy draft. RV, MC, PP, JPT and PV helped to draft the manuscript and have revised and approved the final manuscript.

Acknowledgements: Not applicable

References

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2021 Feb 4;caac.21660.
2. Canfell K. Towards the global elimination of cervical cancer. Vol. 8, Papillomavirus Research. Elsevier B.V.; 2019. p. 100170.
3. World Health Organization. Guidelines for screening and treatment of precancerous lesions for cervical cancer prevention. 2013;
4. Gravitt PE, Paul P, Katki HA, Vendantham H, Ramakrishna G, Sudula M, et al. Effectiveness of VIA, pap, and HPV DNA testing in a cervical cancer screening program in a Peri-Urban community in Andhra Pradesh, India. PLoS One. 2010;5(10).
5. Bigoni J, Gundar M, Tebeu PM, Bongoe A, Schäfer S, Fokom-Domgue J, et al. Cervical cancer screening in sub-Saharan Africa: A randomized trial of VIA versus cytology for triage of HPV-positive women. Int J Cancer. 2015 Jul 1;137(1):127–34.
6. De Vuyst H, Claeys P, Njiru S, Muchiri L, Steyaert S, De Sutter P, et al. Comparison of pap smear, visual inspection with acetic acid, human papillomavirus DNA-PCR testing and cervicography. Int J Gynecol Obstet. 2005;89(2):120–6.
7. Zuchna C, Hager M, Tringler B, Georgoulopoulos A, Ciresa-Koenig A, Volgger B, et al. Diagnostic accuracy of guided cervical biopsies: A prospective multicenter study comparing the histopathology of simultaneous biopsy and cone specimen. Am J Obstet Gynecol. 2010;203(4):321.e1-321.e6.
8. Massad LS, Jeronimo J, Katki HA, Schiffman M, Antani S, Boardman L, et al. The accuracy of colposcopic grading for detection of high-grade cervical intraepithelial neoplasia. J Low Genit Tract Dis. 2009 Jul;13(3):137–44.
9. Fujita H. AI-based computer-aided diagnosis (AI-CAD): the latest review to read first. Radiol Phys Technol 2020 131. 2020 Jan 2;13(1):6–19.
10. Xu T, Kim E, Huang X. Adjustable adaboost classifier and pyramid features for image-based cervical cancer diagnosis. In: Proceedings - International Symposium on Biomedical Imaging. IEEE Computer Society; 2015. p. 281–5.
11. Xu T, Xin C, Long LR, Antani S, Xue Z, Kim E, et al. A new image data set and benchmark for cervical dysplasia classification evaluation. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Springer Verlag; 2015. p. 26–35.
12. Xu T, Zhang H, Xin C, Kim E, Long LR, Xue Z, et al. Multi-feature based benchmark for cervical dysplasia classification evaluation. Pattern Recognit. 2017 Mar 1;63:468–75.

13. 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. JNCI J Natl Cancer Inst. 2019 Sep 1;111(9):923–32.
14. Alyafeai Z, Ghouti L. A fully-automated deep learning pipeline for cervical cancer classification. Expert Syst Appl. 2020 Mar 1;141:112951.
15. Cho B-J, Choi YJ, Lee M-J, Kim JH, Son G-H, Park S-H, et al. Classification of cervical neoplasms on colposcopic photography using deep learning. Sci Reports 2020 101. 2020 Aug 12;10(1):1–10.
16. Xue Z, Novetsky AP, Einstein MH, Marcus JZ, Befano B, Guo P, et al. A demonstration of automated visual evaluation of cervical images taken with a smartphone camera. Int J Cancer. 2020 Nov 19;147(9):2416–23.
17. Peng G, Dong H, Liang T, Li L, Liu J. Diagnosis of cervical precancerous lesions based on multimodal feature changes. Comput Biol Med. 2021 Mar 1;130.
18. Viñals R, Vassilakos P, Rad MS, Undurraga M, Petignat P, Thiran JP. Using dynamic features for automatic cervical precancer detection. Diagnostics. 2021;11(4).
19. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021 Mar 29;372.
20. A.W. Harzing. Publish or Perish. 2007.
21. Whiting PF, Rutjes AWS, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. Quadas-2: A revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med. 2011;155(8):529–36.
22. Krizhevsky A, Sutskever I, Hinton GE. ImageNet Classification with Deep Convolutional Neural Networks.
23. Ren S, He K, Girshick R, Sun J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Trans Pattern Anal Mach Intell. 2015 Jun 4;39(6):1137–49.
24. Szegedy C, Ioffe S, Vanhoucke V, Alemi A. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. 31st AAAI Conf Artif Intell AAAI 2017. 2016 Feb 23;4278–84.
25. Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition. 3rd Int Conf Learn Represent ICLR 2015 - Conf Track Proc. 2014 Sep 4;
26. He K, Zhang X, Ren S, Sun J. Identity Mappings in Deep Residual Networks. In: Leibe B, Matas J, Sebe N, Welling M, editors. Computer Vision – ECCV 2016. Cham: Springer International Publishing; 2016. p. 630–45.
27. Hilal Z, Tempfer CB, Burgard L, Rehman S, Rezniczek GA. How long is too long? Application of acetic acid during colposcopy: a prospective study. Am J Obstet Gynecol. 2020 Jul 1;223(1):101.e1-101.e8.

Figures

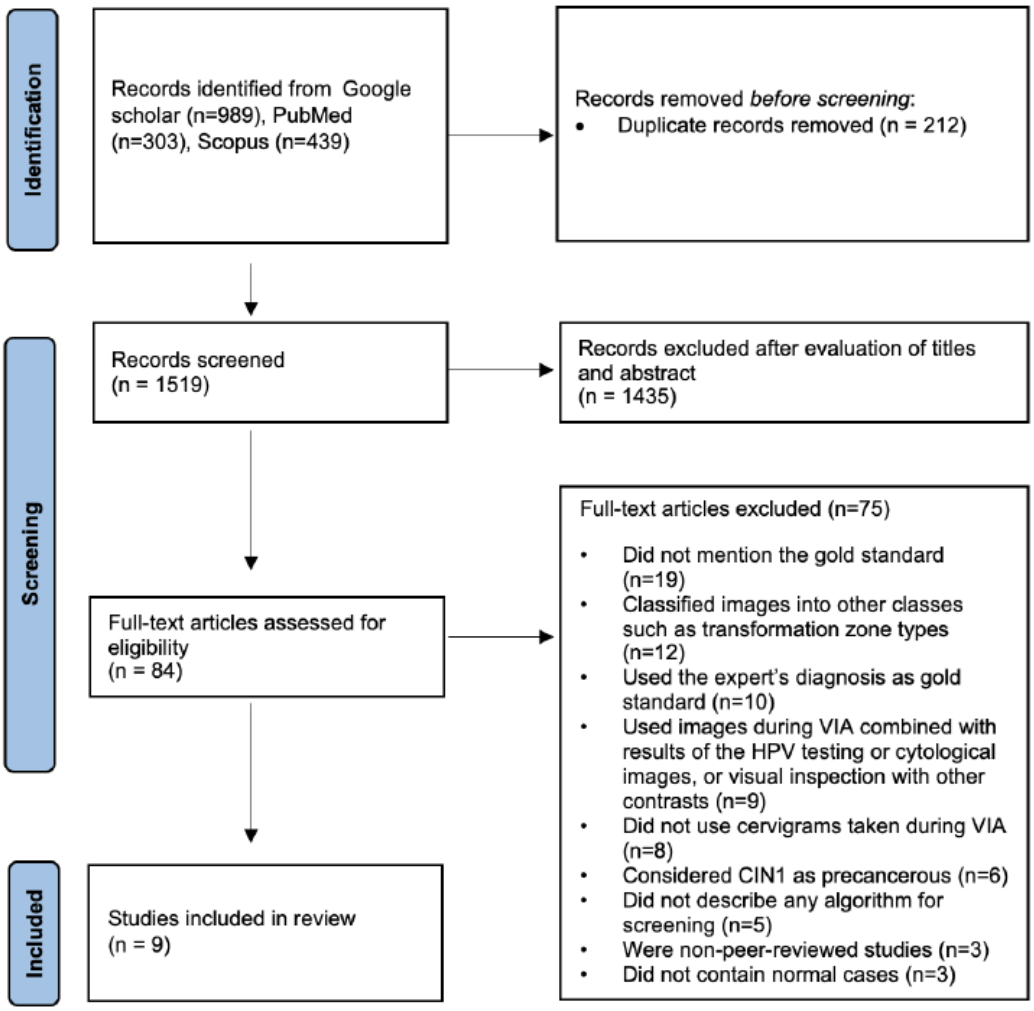

Figure 1

PRISMA flowchart of study selection. Adapted from (19).

Figure 2

Sensitivity and specificity plots AI: artificial Intelligence; ANN: artificial neural networks; CIN: cervical intraepithelial neoplasia; CNN: convolutional neural network; ROC: receiver operating characteristic

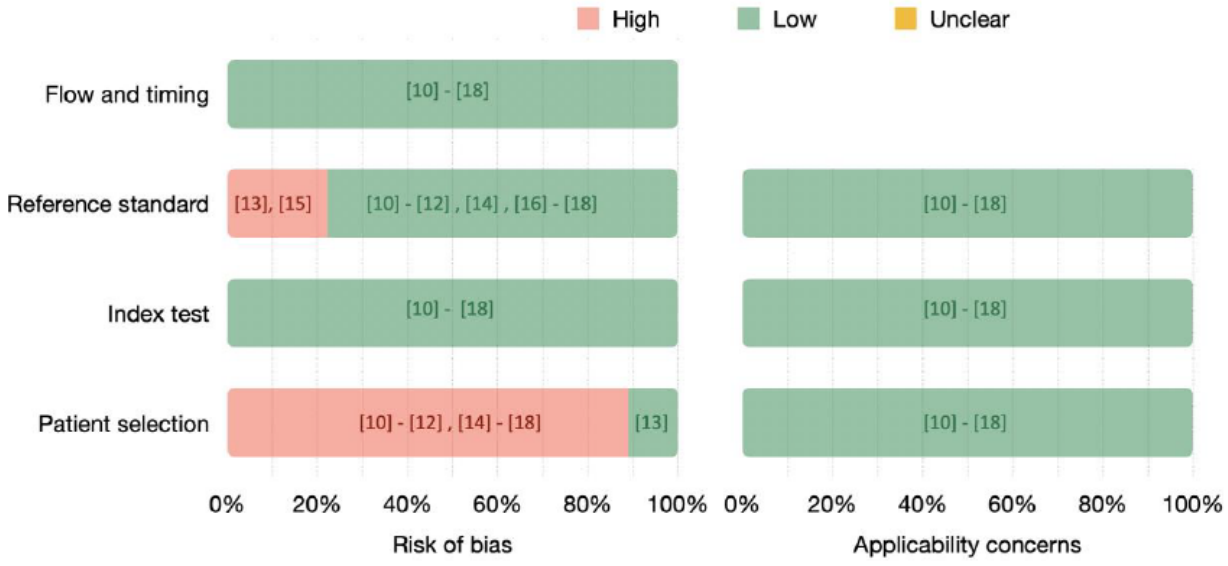

Figure 3

QUADAS-2 assessment of included studies. Risk of bias (left) and applicability concerns (right) are represented in percentages and the references and of the respective studies are indicated.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1.docx
- Additionalfile2.docx