A review of Image Analysis and Machine Learning Techniques for Automated Cervical Cancer Screening from pap-smear images

Wasswa William[1], Annabella Habinka Basaza-Ejiri (PhD)[2] , Andrew Ware (PhD)[3]

[1]*Department of Biomedical Sciences and Engineering, Mbarara University of Science and Technology, Uganda*
[2]*College of Computing and Engineering, St. Augustine International University, Uganda*
[3]*Faculty of Computing, Engineering and Science, University of South Wales, UK*

Abstract.

Background and Objective: Early diagnosis and classification of a cancer type can help facilitate the subsequent clinical management of the patient. Cervical cancer ranks as the fourth most prevalent cancer affecting women worldwide and its early detection provides the opportunity to help save life. To that end, automated diagnosis and classification of cervical cancer from pap-smear images has become a necessity as it enables accurate, reliable and timely analysis of the condition's progress. This paper presents an overview of the state of the art as articulated in several prominent recent publications focusing on automated detection of cervical cancer from pap-smear images.

Methods: The survey reviews publications on applications of image analysis and machine learning in automated diagnosis and classification of cervical cancer from pap-smear images spanning 15 years. The survey reviews 30 journal papers obtained electronically through four scientific databases: Google Scholar, Scopus, IEEE and Science Direct searched using three sets of keywords: segmentation, classification, cervical Cancer; mmedical imaging, machine learning, pap-smear; automated system, classification, pap-smear.

Results: Most of the existing algorithms offer the accuracy of nearly 93.78% on an open pap-smear data set, segmented using CHAMP digital image software. K-Nearest-Neighbors and Support Vector Machines algorithms have been reported to be excellent classifiers for cervical images with accuracies of over 99.27% and 98.5% respectively when applied to a 2-class classification problem (normal or abnormal).

Conclusion: The reviewed papers indicate that there are still weaknesses in the available techniques which result in low accuracy of classification in some classes of cells. Moreover, most of the existing algorithms work either on single or multiple cervical smear images. This accuracy can be improved by using varying various parameters such as the features to be extracted, improvement in noise removal, using hybrid segmentation and classification techniques like construction of multi-level classifiers. Combining K-Nearest-Neighbors algorithm with other algorithm(s) such as Support Vector Machines, pixel level classifications and including statistical shape models can improve performance. Further, most of the developed classifiers are developed and tested on accurately segmented images using commercially available software such as CHAMP software. There is thus a deficit of evidence that these algorithms will work in clinical settings found in developing countries (where 85% of cervical cancer incidences occur) that lack sufficient trained cytologists and the funds to buy the commercial segmentation software.

Keywords: *Machine Learning, Medical Imaging, Segmentation*

1.0 Introduction

Globally, cervical cancer ranks as the fourth most prevalent cancer affecting women worldwide after breast, colorectal, and lung cancers [1], with 527,624 women diagnosed with cervical cancer and 265,672 dying from the disease every year worldwide [2]. In sub-Saharan Africa, 34.8 new cases of cervical cancer are diagnosed per 100,000 women annually, and 22.5 per 100,000 women die from the disease [1], with over 80% of cervical cancer detected in its late stages. Over 85 per cent of cervical cancer cases occur in less developed countries of which the highest incidences are in Africa, with Uganda being ranked 14th among the countries with the highest incidences of cervical cancer [3]. Over 65% of those diagnosed with the disease in Uganda die from it [4]. This is attributed to lack of awareness of the disease and limited access to health services**. In contrast, developed countries have strategies to enable reliable and effective screening methods and thus pre-cancerous lesions are detected and treated at an earlier stage [5]. As a strategy for reducing the occurrences of cervical cancer in Sub-Saharan Africa, the

World Health Organization recommends screening and vaccination throughout the sub-Saharan African region so as to help achieve the UN Sustainable Development Goal 3 of ensuring healthy lives and promoting well-being for all [6]. The government of Uganda through the Ministry of Health launched its strategic plan for cervical cancer prevention and control in 2010 with the target of screening and vaccinating at least 80% of the eligible population by 2015 [7].

Cervical cancer can be prevented if effective screening programmes are in place and this can lead to reduced morbidity and mortality [8]. The success of screening has been reported to depend on a number of factors including: access to facilities, quality of screening tests, adequacy of follow-up, and diagnosis and treatment of lesions detected [9]. Cervical cancer screening services are very low in low and middle income countries due to the presence of only a few trained and skilled health workers, and lack of healthcare resources to sustain screening programmes [10]. It is estimated that only a small percentage (5%-27%) of women in sub-Saharan Africa report having received cervical cancer

screening [11]. This is even lower in the East African region where cervical cancer age-standardized incidence rates are highest due to inadequate screening programs [12].

This paper reviews several image analysis and machine learning techniques proposed by different researchers for automated cervical cancer screening from pap-smear images. The paper is organised as follows: The rest of section I gives information about pap-smear images. Section 2 outlines the methodology used to obtain the reviewed papers. Section 3 presents a detailed review of medical imaging and machine learning technical proposed by several selected researchers. Observations and Discussions are presented in Section 4 and, finally, conclusions and future research are presented in section 5.

1.1 The Papanicolaou test

The pap-test is a manual screening procedure which is used to detect pre-cancerous changes in cervical cells on the basis of color and shape properties of cell nuclei and cytoplasm regions [13]. The test is the commonest technique used for early screening and diagnosis of cervical cancer. Pap-smears have helped to reduce the mortality rate of cervical cancer by between 50% and 70% in developing countries [14]. Samples are observed under a microscope in order to detect any unusual developments indicating any precancerous and potentially precancerous changes. Examining the cell images for abnormalities in the cervix provides grounds for provision of prompt action and thus reducing incidence and deaths from cervical cancer. Pap-smear tests, if done with a regular screening programs and proper follow-up, can reduce cervical cancer mortality by up to 80% [14].

The manual analysis of the pap-smear images is time consuming, laborious and error prone. During each patient screening, hundreds of sub-images (as shown in Figure 1) have to be looked at by a trained cytologist using a microscope. This makes the screening process very tedious, labor intensive and erroneous [15]. Furthermore, human visual grading for microscopic biopsy images is very subjective and inconsistent due to inter-and intra-observer variations and monotony of the task as cytotechnologists usually screen a large amount of data on a daily basis.

The pap-smear slide may contain a single cervical cell, clusters of cervical cells, cervical cells together with other cells (for example white blood cells, red blood cells), together with miscellaneous debris or even bacteria. The appearance of the cells in the pap-smear image usually depends on how the pap-smear was obtained from the cervix, stained and digitized. Staining makes the cells in the images appear in different colors and the size of the cells in the image depend on the magnification of the lens used during image acquisition, while the type of camera used for digitization determines the quality of the pap-smear images.
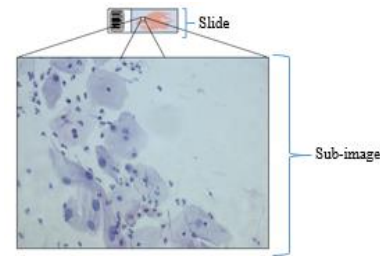

Figure 1: A typical pap-smear image (slide) and a high resolution field of view (sub-image). Approximately 10,000 sub-images are needed to cover the whole slide.

1.2 Features of cervical cells in a pap-smear image analysis

The cervical cells are divided into seven classes (superficial squamous epithelial, intermediate squamous epithelial, columnar epithelial, mild dysplasia, moderate dysplasia, severe dysplasia, and carcinoma-in-situ) depending on the cell nucleus [16]. The cell nucleus is usually used for cervical cancer screening and classification as it contributes to the cell changes when a cell has been affected and its properties such as size, shape and intensity are usually compared during cell classification as normal or abnormal.

Abnormal cervical cells are called dysplastic cells, which have four classes. The first class is called the mild dysplastic. In mild dysplastic, the nucleus becomes larger and brighter than normal; however, a number of mild dysplastic cells will disappear without becoming malignant [19]. The second class is called moderate dysplastic where the nucleus becomes much larger and darker [19]. The third class is known as the severe dysplastic where the nucleus and cytoplasm both change their size and texture [18]. A number of severe dysplastic cells turn into malignant cells. In severe dysplastic cells, the nucleus becomes larger, deformed and darker but the cytoplasm is usually darker and smaller. The other form of abnormal cells is Carcinoma-in-situ [20]; which possess more serious problem than severe dysplastic.

Normal cells taken from the cervix usually contain cells from the columnar epithelium and the squamous epithelium [20]. The squamous epithelium cells form at the basal layer (the deepest layer of the epidermis), move to the intermediate layer and finally to the superficial layer (a thin layer of subcutaneous connective tissue that lies between the dermis of the skin and the deep cervical fascia). The cells change shape and color as they move through the different layers. The columnar epithelium contains reserve and columnar cells. The nucleus is usually located at the bottom of the cytoplasm: in normal columnar cells the nucleus appears larger when viewed from the top while the cytoplasm looks larger when viewed from the sides.

A number of important specific cell features (Table 1) are used for cervical cell image analysis. These features are categorized as: Size (cell area, nucleus area, cytoplasm area etc.); Shape (nucleus roundness, cytoplasm roundness etc.); Ratio (nucleus/cytoplasm ratio, Percentage of empty cells etc.); Topology (Distribution of nucleus, nucleus position etc.); Color intensity (cell, nucleus, cytoplasm brightness etc.); and Texture (Multi-nucleus cells etc.) [19]. These features are an important input attribute for any cervical cancer classification system.

Table 1: Some of the Cell Features used for cervical cancer classification [19]

Feature	Cervical Cancer Class			
	Normal	Degree of Dysplasia		
		Mid	Moderate	Severe
Nucleus Area(μm^2)	20-50	50+	50+	50+
Nucleus Intensity	dark	light	dark	Dark
Cytoplasm Intensity	light	light	dark	Dark
Nucleus/Cytoplasm-Ratio	1-2%	10-20%	20-50%	50% +

1.3 Automated pap-smear analysis.

The aim of automated pap-smear analysis is to segment and then classify cervical cells in the pap-smear images as either normal or abnormal [19]. Given the progress in the development of such techniques as medical imaging and machine learning, such automated analysis is now a reality, helping to reduce the time spent and increase the accuracy of cytologists performing slide examination during the pap screening process.

Medical image analysis involves the techniques and processes to obtain detailed information from medical images for clinical analysis and medical interventions [19]. Machine learning (ML) techniques are a branch of artificial intelligence that relates to the problem of learning from data samples. Machine learning employs a variety of statistical, probabilistic and optimization techniques that allow computers to "learn" from past examples and to detect hard-to-discern patterns from large, noisy or complex data sets [20]. Machine learning and medical imaging techniques make it possible to automatically analyze pap-smear images and make the screening process faster and more reliable, as proposed by various papers presented in the literature review.

1.3.1 Stages and techniques used in automated pap-smear analysis

A typical automated pap-smear analysis system consists of five stages, namely: Image acquisition; pre-processing; segmentation; feature extraction; and classification.

1) Image acquisition involves the tools and techniques used to obtain digital images from pap-smear slides.
2) Pre-processing involves image enhancement, background extraction and definition of regions of interest in the images.
3) Segmentation aids in extracting regions of interest in a pap-smear image. For cytological images, the main focus is generally to isolate the cell nucleus since the malignant or abnormal characteristics are most prominent there.
4) Feature extraction involves describing properties of the selected regions of interest. The two main approaches to describing these properties can be classified as the structure approach and the texture approach.
5) Classification seeks to to determine whether or not a sample contains any evidence of cancer (or precancerous lesions). This is the main aim of the analysis of pap-smears.

2.0 Methods.

The survey reviews publications on applications of image analysis and machine learning in automated diagnosis and classification of cervical cancer from pap-smear images spanning 15 years. The survey reviews 30 journal papers obtained electronically through four scientific databases: Google Scholar, Scopus, IEEE and Science Direct searched using *three sets* of keywords:

Set 1: Segmentation and Classification; and Cervical Cancer.

Set 2: Medical Imaging and Machine Learning; and Pap-smear.

Set 3: Automated System and Classification; and Pap-smear.

From Google Scholar, *Set 1*, *Set 2* and *Set 3* Keywords produced 16,200, 2,090 and 5,380 search results respectively. IEE produced 25, 7 and 5 search results from *set 1, set 2* and *set 3* keywords respectively. 87, 65 and 32 search results were produced from Scopus using *set 1, set 2* and *set 3* key words respectively. Finally 210, 20 and 3 search results were obtained from Science Direct data base using *set 1, set 2* and *set 3* keywords respectively.

3.0 Results

This section documents the findings of a literature review, of publications spanning a fifteen-year period, relating to the applications of image analysis and machine learning in automated diagnosis and classification of cervical cancer from pap-smear images obtained using the search criteria presented in the methods section.

Out of the 24,124 search results obtained from the four scientific databases considered using the three sets of key words; only 30 papers were considered for this review article. Papers that have been frequently cited have been found more relevant for this review. Papers presenting prediction of prevalence and recurrence of cervical cancer using machine learning techniques were excluded. Furthermore, papers were the accuracy or sensitivity or specificity or validity of the algorithm has not been reported have not been considered in this review paper.

3.1 Survey of Existing Literature on automated pap-smear analysis

Y. Song et al. (2017) [21] proposed a learning-based method with shape models to segment individual cell in pap-smear images. The splitting of the cells was defined as a discrete labeling task with a suitable cost function. The labeling results were then fed into a dynamic multi-template deformation model for further boundary refinement. An evaluation carried out using two different datasets demonstrated the superiority of the proposed method over the state-of-the-art methods in terms of segmentation accuracy.

B. Ashok et al. (2016) [22] compared feature selection methods for diagnosis of cervical cancer using a Support Vector Machine (SVM) classifier. Image segmentation was performed using thresholding. Feature selection was achieved using mutual information, sequential forward search, sequential floating forward search, and random subset feature selection methods. Accuracy of 98.5%, sensitivity of 98% and specificity of 97.5% were obtained using the sequential

floating forward selection method, which was higher than the other methods.

H. Lee et al. (2016) [23] proposed an automatic segmentation method for multiple overlapping cervical cells in microscopic images using superpixel partitioning and cell-wise contour refinement. The cells are detected using superpixel generation and triangle thresholding. The nuclei are extracted using local thresholding and cytoplasm by superpixel partitioning. The method showed competitive performances compared to other methods.

J. Su et al. (2016) [24] proposed a method for automatic detection of cervical cancer from pap-smear images using a two-level cascade integration system of two classifiers. The results showed that the recognition rates for abnormal cervical cells were 92.7% and 93.2%, respectively, when C4.5 classifier or LR (logical regression) classifier were used individually; while the recognition rate was significantly higher (95.6%) when the two-level cascade integrated classifier system was used.

M. Sharma et al. (2016) [25] used K-Nearest-Neighbors (KNN) method to classify the stage of cervical cancer from pap-smear images. The classification accuracy of 82.9% with 5-Fold cross validation was achieved.

R. Kumar et al. (2015) [26] proposed a framework for automated detection and classification of cervical cancer from microscopic biopsy images using biologically interpretable features. K-means is used for image segmentation and K-nearest neighborhood is used for cervical cancer classification. The performance measures for accuracy, specificity and sensitivity of 92%, 94% and 81% were obtained.

Y.song et al. (2015) [27] proposed a multiscale convolutional network (MSCN) and graph-partitioning-based method for segmentation of cervical cytoplasm and nuclei. Deep learning via the MSCN was used to extract scale invariant features, and then segment regions centered at each pixel. Experimental results demonstrate that the proposed approach delivers promising results.

T. Chankong et al. (2014) [28] presented a method for automatic cervical cancer cell segmentation and classification using fuzzy C-means (FCM) clustering technique. Validation with Artificial Neural Networks (ANN) yielded accuracies of 93.78% and 99.27% for the 7-class and 2-class problems, respectively.

Y. Song et al. (2014) [29] applied a super pixel and convolution neural network (CNN) based segmentation method to cervical cancer cells. They also explored the use of Deep learning based on CNN for region of interest detection. Experimental results of 94.50% were achieved for nucleus region detection and a precision of 0.91±0.02 and a recall of 0.87±0.001 were achieved for nucleus cell segmentation.

J. Talukdar et al. (2013) [30] presented a fuzzy clustering based image segmentation of pap-smear images of cervical cancer cells using Fuzzy C-Means (FCM) Algorithm. Two random numbers were used to form the membership matrix for each pixel to guide clustering. Promising results were obtained using the pixel level segmentation.

Z. Lu et al. (2013) [31] presented an algorithm for the segmentation of cytoplasm and nuclei from clumps of overlapping cervical cells. Their approach addresses challenges involved in delineating cells with severe overlap by utilizing a joint optimization of multiple level set functions. Their quantitative assessment demonstrates that the methodology can successfully segment clumps of up to 10 cells, provided the overlap between pairs of cells is < 0.2 mm.

Z. Lu et al. (2013) [32] presented a joint level set optimization method for automated nucleus and cytoplasm segmentation of cervical cells using scene segmentation and unsupervised classification. The method obtained a Jaccard index of > 0.8 with a near zero false negative rate.

A. Genctav et al. (2012) [33] proposed an unsupervised approach for the segmentation and classification of cervical cells. The approach involves automatic thresholding to separate the cell regions from the background. A multi-scale hierarchical segmentation algorithm was used to partition the regions of interest based on homogeneity and circularity. Finally, a binary classifier was used to separate the nuclei from cytoplasm. Performance evaluation using two data sets showed the effectiveness of the proposed approach with an accuracy of 96.71%.

A.Kale et al. (2012) [34] presented a nucleus segmentation technique which determines a segmentation threshold based on the stability of the perimeter of the cell. Cytoplasm and nucleus are separated by clustering. A minimum Mahalanobis distance classifier was used to compare results. The technique achieved an accuracy of 90.0% for two class problems classification.

C. Bergmeir et al. (2012) [35] implemented an algorithm for segmenting the nuclei from pap-smear images. The algorithm localizes cell nuclei using a voting scheme and prior shape knowledge by means of an elastic segmentation algorithm. Edges are extracted with a Canny edge detection algorithm and a randomized Hough transform to find candidate nuclei, which are then processed by a level set algorithm. Experiments showed promising results.

P.Pai et al. (2012) [36] presented a nucleus and cytoplast contour detector (NCC) for cytoplast and nucleus segmentation in pap-smear images using maximal gray-level-gradient-difference (MGLGD) method. Adaptable threshold decision (ATD) method was utilized to separate the cells in the pap-smear images. Results showed that the proposed method is superior to the gradient vector flow-active contour model (GVF-ACM) and the edge enhancement nucleus and cytoplast contour (ENNCC) detector, in segmenting the cytoplast and nucleus of a cell.

I. Muhimmah et al. (2012) [37] presented a novel method for nuclei segmentation using morphological operation and watershed transformation. The method produced promising

results when evaluated with respect to its nuclei area and its shape-similarity in comparison to the pathologist truth.

K. Li et al. (2012) [38] proposed a Radiating Gradient Vector Flow (RGVF) snake algorithm to extract nucleus and cytoplasm from single cervical cell image. Special k- means algorithm was used to cluster the image into areas of nucleus, cytoplasm and background. Experiments performed on the Herlev dataset showed that the proposed algorithm is effective.

M. Sreedevi et al. (2012) [39] presented an algorithm based on iterative thresholding method for segmentation of pap-smear images and classification of cervical cells as normal or abnormal based on the area parameter of the nucleus. The features of the nucleus were extracted using region properties, and cells were classified as normal if nucleus area was less than 1635mm and classified as abnormal otherwise. A sensitivity of 100% and specificity of 90% was achieved.

H. Kong et al. (2011) [40] proposed an integrated framework consisting of a novel supervised cell-image segmentation algorithm and a touching-cell splitting method. For the segmentation, the color-texture was extracted at the local neighbourhood of each pixel using a local Fourier transform (LFT). The boundaries of touching-cell clumps were smoothed out by Fourier shape descriptor. The pipeline was validated against pathological images giving an error rate of 5.25% per image in terms of under-splitting, over-splitting, and encroachment errors.

L. Zhang et al. (2011) [41] presented a nuclei segmentation algorithm consisting of three main components: pre-processing, binarization and segmentation. In pre-processing, HSV color space was used for enhancing the contrast between nuclei and cytoplast. An adaptive threshold method was used during binarization to separate the nuclei pixels from background pixels. For segmentation, a concave point based overlapped nuclei segmentation algorithm was utilized.

M. Plissiti et al. (2011) [42] presented an automated method for the detection and boundary determination of cells nuclei in pap-smear images. The detection of the candidate nuclei was based on a morphological image reconstruction process and the segmentation of the nuclei boundaries was accomplished with the application of the watershed transform. The method was evaluated on a data set of 90 pap-smear images. Comparisons with the segmentation results of a gradient vector flow deformable (GVF) model and a region based active contour model (ACM) indicate that the method produces more accurate nuclei boundaries that are closer to the ground truth.

S.Sulaiman et al. (2010) [43] proposed a segmentation method for delineating the overlapping cells in pap-smear images. A seed based region growing algorithm was utilized to detect and segment overlapping cells. A pseudo coloring technique was used to delineate the cytoplasm and nucleus.

C. Lin et al. (2009) [44] proposed a method for detection and segmentation of cytoplasm and nucleus from pap-smear images. The approach used a Gaussian filter for noise elimination and a two-group object enhancement technique to enhance the gradients of the edges of the cytoplasm and nucleus while suppressing the noise gradients. Performance was compared with seed region growing feature extraction and level set method and showed promising results.

Y. Marinakis et al. (2009) [45] proposed a meta-heuristic algorithm to classify cervical cells from pap-smear images using a genetic algorithm scheme combined with a number of nearest neighbor based classifiers. Results showed that classification accuracy generally outperforms other previously applied intelligent approaches with accuracy of about 89%.

M. Plissiti et al. (2008) [46] presented an automated method for cell nuclei detection. Fuzzy C-means algorithm was used for segmentation and clustering. The proposed method was evaluated on a data set consisting of 3,085 cells of pap-smear images and showed promising results.

S. Yang et al. (2008) [47] presented an edge enhancement nucleus and cytoplast contour (EENCC) detector to enable cutting the nucleus and cytoplast from a cervical smear cell image for automated cervical cancer diagnosis. A trim-meaning filter was used to effectively remove impulse and Gaussian noise.

N. Ampazis et al. (2004) [16] proposed an algorithm for cervical cancer screening using efficient second order neural network. The classification algorithms used were the LMAM (Levenberg Marquardt with Adaptive Momentum) and OLMAM (Optimized Levenberg-Marquardt with Adaptive Momentum) which resulted into an overall accuracy of 80.7%.

J. Zhang et al. (2004) [49] presented a novel feature screening method by deriving relevance measures from the decision boundary of Support Vector Machine using pixel-level classification for feature selection. Comparative experiments with other algorithms showed significant improvements on pixel-level classification accuracy using the new set of derived features

4.0 Observations and Discussion

The overall aim of the cervical cells preprocessing, segmentation and classification algorithms summarized in this paper is automated diagnosis and classification of cervical cancer from pap-smear images [16, 21-49]. To that end, the reviewed papers document adaptations to the various stages of the medical image analysis pipeline which include: image acquisition, pre-processing, segmentation, feature extraction and classification [50].

The review of relevant literature has highlighted that some techniques are more frequently used than others; with filtering, thresholding and KNN being the most frequently used techniques for preprocessing, segmentation and classification of pap-smear images respectively. It has also been observed that the superiority of the results of a classification algorithm over the other greatly depends on a number of factors that include: the accuracy of the segmentation, pre-processing and the type of datasets used. Table 1 shows a summary of the techniques used in the papers reviewed.

Table 1: A summary of the techniques documented in the papers reviewed.

Image Acquisition	Pre-processing	Segmentation	Classification
-Most of the papers reviewed use open datasets. The commonly being: - LCH - ERUDIT and - Herlev datasets -A few authors report to have obtained pap-smear from hospitals with support of skilled cytotechnicians	*Filters:* The commonly used being the Median and the Gaussian Filter	- Adaptive histogram equalization - Thresholding - Hough Transformation - Superpixel based Markov Random Field - Seed Based Region Growing - Watershed segmentation - Shape Prior - Edge detection	- K-means - Artificial Neural Networks - K-Nearest Neighbor - Fuzzy C-Mean - Support Vector Machines - Bayesian Networks - Decision Trees

The review has identified that there are still some weak points with regard to the techniques reviewed; these weaknesses include low accuracy of classification in some classes of cells. Furthermore, the algorithms documented work either on single cervical cell images or multiple cervical smear images; hence algorithms that can be used on both single and multiple cell images at the same time should be explored as cells in pap-smear images usually appear as overlapping cells.

Most of the existing algorithms result in accuracy of nearly 93.78% (which is still low) on an open pre-processed pap-smear data set (Herlev dataset images) located at http://mde-lab.aegean.gr/downloads. The reported accuracy can be improved up to the higher level by varying various parameters of the algorithms such as the features to be extracted, improvement in noise removal methods, using hybrid segmentation and classification techniques.

As reported earlier, some segmentation and classification algorithms are more frequently used than others; a situation that might have arisen due to the various advantages of one technique over the other (Table 2 and Table 3).

5.0 Conclusions

The paper reviews a number of prominent recent publications relating to the automated diagnosis and classification of cervical cancer from pap-smear images. This review should help researchers in the field to see the challenges associated with some of the techniques documented, and provided a good basis for designing and developing new algorithms or improve existing ones.

KNN algorithm has been reported to be an excellent classifier for cervical images, however combining KNN algorithm with other algorithm (s) like Support Vector Machines (SVMs), pixel level classifications and including statistical shape models can improve performance. Furthermore application of multi-level segmentation can improve performance of a classifier (s).

Most of the algorithms have been cross validated with the training and test datasets provided from the online cervical image datasets like Herlev (http://mde-lab.aegean.gr/downloads). Other new images from hospitals should be used for testing the developed classifiers by a trained cytologist and results reported. Finally most of the developed classifies are developed and tested on accurately pre-processed images segmented using commercially available segmentation softwares like CHAMP digital image software. There is thus a deficit of evidence that these algorithms will work in clinical settings found in developing countries (where 85% of cervical cancer incidences occur) that lack sufficient trained cytologists and the funds to buy the commercial segmentation softwares.

Declarations

☐ Competing interests

☐ Acknowledgements

Table 2: Advantages and Limitations of the Segmentation classification Techniques reviewed in this paper

	Technique	Advantages	Limitations
1	**Segmentation**		
	Water immersion algorithms	Closed boundaries can be detected	Not good for overlapping cells
	Active contours	Boundaries can easily be detected	Energy minimization is time consuming
	Hough transform	Easily detects similarity in shapes	Detects only round single shapes
	Fuzzy logic	Robust and handles uncertainty in data	Rigid rules and not flexible
	Seed based region growing algorithms	Detects edges	Edge detection is subject to user's defined parameters
	Genetic algorithms	Efficiently searches for a good solution	Slow
	K-means clustering	Threshold values detected automatically	Cannot separate overall aping cells.
2	**Classification**		
	Artificial Neural Networks	Tolerant to noise, Can use more than one instance to classify instances	Over fitting in case of many attributes and complex algorithm
	Bayesian Networks	Easy to understand, based on statistical inference	Assumes normal distribution on numeric attributes, assumes attributes are statistically independent
	Support Vector Machines	Easy to control decision tree and over fitting is unlikely to occur.	Difficult to obtain optimal parameters for nonlinear data and training is slow.
	Decision Tree	Easy to understand	Error in training set can lead to wrong final decisions.
	Genetic algorithms	Finds a good solution and good for optimization	Usually finds local optima that makes it not the most efficient
	k-Nearest Neighbour	Robust, fast and tolerant to noise	Complex as attributes increase in number and assumes that instances of same attributes are similar. This may not be true all the times.

Table 2: Review of automated cervical cancer screening papers from pap-smear images

Year	Author	Paper	Datasets	Features	Preprocessing	Segmentation	Classification	Results	Limitations
2016	M. Sharma et al	Classification of Clinical Dataset of Cervical Cancer using KNN	Single cells Data sets from Fortis Hospital Mohali, Punjab (India)	7 morphological features	Gaussian filter and Histogram equalization	Features normalized using min-max and Edge Detection methods	K-Nearest Neighbor method	Classification accuracy with 82.9% of maximum performance with 5 Fold cross validation	Validated with few futures and only applicable to single cells
2016	J. Su et at.	Automatic Detection of Cervical Cancer Cells by a Two-Level Cascade Classification System	5000 epithelial cells from 120 liquid based cytology slides	28 features including 20 morphological and 8 texture features	Histogram equalization and median filter	Adaptive threshold segmentation	A two- level integration system of two classifiers (C4.5 and Logical Regression)	Recognition rates of 92.7% and 93.2%, when C4.5 or LR classifier was used individually and 95.6% for the two-level cascade integrated classifier	Only looks at Normal or abnormal (2-class problem)
2016	B. Ashok et al.	Comparison of Feature selection methods for diagnosis of cervical cancer using SVM classifier,	150 images of pap-smear test are collected from Rajah Muthiah Medical College	14 Texture and 30 Shape features	Filters, Image resizing and Gray scale image conversation.	Multi-Thresholding method	SVM classifier	Accuracy 98.5%, sensitivity 98% and specificity 97.5% obtained from SFFS which is higher than other methods	Small validation data set
2015	R. Kumar et al.	Detection and Classification of Cancer from Microscopic Biopsy Images Using Clinically Significant Features	2828 histology images from the histology image dataset (histologyDS2828)	125 morphologic Features	Contrast limited adaptive histogram equalization	K-means segmentation algorithm	K- NN, fuzzy KNN, SVM and random forest based classifiers	Accuracy, Specificity and Sensitivity of 92%, 94% and 81% were obtained	KNN has been compared to only fuzzy KNN, Random forest and SVM.
2014	T. Chan kong et al	Automatic cervical cell segmentation and classification in pap-smears	Herlev dataset	Cytoplasm, nucleus and background	Pre-processed dataset	Patch-based fuzzy C-means and FCM clustering method	FCM algorithm	Accuracies of 93.78% and 99.27% for the 7-class and 2-class problems.	Not applicable for cervical cell image with multiple cells
2013	J. Talukdar et al.	Fuzzy Clustering Based Image Segmentation of pap-smear Images of Cervical Cancer Cell Using FCM Algorithm	Color image	Morphometric, Densitometric, Colorimetric and textural feature	Adaptive histogram Equalization with Otsu's method	General and random numbers generated based on Chaos theory corresponding to R, G and B value.	Pixel-level classification and shape analysis	Preserves the color of the images and chance of data loss is minimal	Single image used to validate the algorithm

Year	Author	Title	Dataset	Features	Pre-processing	Segmentation	Classification	Results	Limitations
2013	Z. Lu et al.	Automated Nucleus and Cytoplasm Segmentation of Overlapping Cervical Cells	A database of 18 synthetically constructed images	Features of the nucleus, Cytoplasm and background	-	Scene segmentation	Unsupervised classification	A Jac-card index of > 0.8 with a near zero false negative rate	Weak transitions between the cytoplasm and background.
2012	M. Sreedevi et al.	Pap-smear Image based Detection of Cervical Cancer,	Herlev University Hospital Dataset	Features of the nucleus such as area.	Pre-processed dataset	Iterative Thresholding method	Based on area parameter of the nucleus	A sensitivity of 100% and specificity of 90% was achieved.	Only looks at Normal or abnormal (2-class problem)
2012	A. Genctav et al	Unsupervised segmentation and classification of cervical cell images	Herlev and Hacettepe pap-smear dataset	Features of the nucleus such as area.	Pre-processed dataset	Two stage segmentation. Multi scale hierarchical and automatic threshold segmentation	5 classifiers. DT Bayesian, SVM and combination of the three using product and sum of individual posterior probabilities.	Accurate segmentation and classification of cervical cell images having inconsistent staining, poor contrast, and overlapping cells	Using accurate cytoplasm features can improve performance. Using already pre-processed images.
2009	Y. Marinakis et al	Pap-smear diagnosis using a hybrid intelligent scheme focusing on genetic algorithm based feature selection and nearest neighbor classification	Herlev University Hospital Dataset	20 numerical features	Pre-processed dataset	Genetic algorithm	Metaheuristic algorithm and nearest neighbor based classifiers	Results significantly improved compared to the results when only the 1-nearest neighbor is used for the classification	2 class problem
2004	N. Ampazis et al.	Pap-Smear Classification Using Efficient Second Order Neural Network Training Algorithms	Herlev University Hospital	20 morphological features	Pre-processed dataset	-	LMAM and OLMAM Algorithms	Classification accuracy for the two class category problem of 98.86% was obtained	Only validated for 2 class classification problem
2002	J. Zhang et al.	SVM Based Feature Screening Applied To Hierarchical Cervical Cancer Detection	Single and Overlapping multispectral pap-smear image database.	68 relevant features derived from the decision boundary of SVM	Background segmentation and Intensity Normalization	Pixel-level analysis and SVM Block wise feature extraction	Pixel-level classification	Significant improvements using pixel level classification	Larger pap-smear image set and a richer image feature space needed to validate the method

References

[1] A. Jemal, M. M. Center, C. DeSantis, and E. M. Ward, "Global patterns of cancer incidence and mortality rates and trends.," *Cancer Epidemiol. Biomarkers Prev.*, vol. 19, no. 8, pp. 1893–907, 2010.

[2] A. Jemal, F. Bray, and J. Ferlay, "Global Cancer Statistics: 2011," *CA Cancer J Clin*, vol. 49, no. 2, pp. 1,33–64, 1999.

[3] L. A. Torre, F. Bray, R. L. Siegel, J. Ferlay, J. Lortet-tieulent, and A. Jemal, "Global Cancer Statistics, 2012," *CA a cancer J. Clin.*, vol. 65, no. 2, pp. 87–108, 2015.

[4] P. Release, "Latest world cancer statistics Global cancer burden rises to 14 . 1 million new cases in 2012 : Marked increase in breast cancers must be addressed.," *Int. Agency Res. Cancer, World Heal. Organ.*, no. December, pp. 2012–2014, 2013.

[5] A. Shariff, J. Kangas, L. P. Coelho, S. Quinn, and R. F. Murphy, "Automated image analysis for high-content screening and analysis," *Journal of Biomolecular Screening*, vol. 15, no. 7. pp. 726–734, 2010.

[6] N. Santesso *et al.*, "World Health Organization Guidelines for treatment of cervical intraepithelial neoplasia 2-3 and screen-and-treat strategies to prevent cervical cancer.," *Int. J. Gynaecol. Obstet.*, vol. 132, no. 3, pp. 252–8, 2016.

[7] C. Nakisige, M. Schwartz, and A. O. Ndira, "Cervical cancer screening and treatment in Uganda," *Gynecologic Oncology Reports*, vol. 20. pp. 37–40, 2017.

[8] C. Schwaiger, M. Aruda, S. Lacoursiere, and R. Rubin, "Current guidelines for cervical cancer screening," *J. Am. Acad. Nurse Pract.*, vol. 24, no. 7, pp. 417–424, 2012.

[9] N. Karjane and D. Chelmow, "New Cervical Cancer Screening Guidelines, Again," *Obstetrics and Gynecology Clinics of North America*, vol. 40, no. 2. pp. 211–223, 2013.

[10] T. Mutyaba, F. A. Mmiro, and E. Weiderpass, "Knowledge, attitudes and practices on cervical cancer screening among the medical workers of Mulago Hospital, Uganda," *BMC Med. Educ.*, vol. 6, 2006.

[11] R. I. Anorlu, "Cervical cancer: the sub-Saharan African perspective," *Reprod. Health Matters*, vol. 16, no. 32, pp. 41–49, 2008.

[12] H. Sancho-Garnier *et al.*, "Overview of Cervical Cancer Screening Practices in the Extended Middle East and North Africa Countries," *Vaccine*, vol. 31, no. 6, pp. G51–G57, 2013.

[13] S. Cox, "Guidelines for Papanicolaou Test Screening and Follow-Up," *J. Midwifery Women's Heal.*, vol. 57, no. 1, pp. 86–89, 2012.

[14] J. Hartikainen, "The Papanicolaou test: its utility and efficacy in cancer detection.," *Contemporary nurse : a journal for the Australian nursing profession*, vol. 11, no. 1. pp. 45–49, 2001.

[15] R. B. Perkins, S. M. Langrish, L. J. Stern, J. Figueroa, and C. J. Simon, "Comparison of visual inspection and Papanicolau (PAP) smears for cervical cancer screening in Honduras: Should PAP smears be abandoned?," *Trop. Med. Int. Heal.*, vol. 12, no. 9, pp. 1018–1025, 2007.

[16] N. Ampazis, G. Dounias, and J. Jantzen, "Pap-Smear Classification Using Efficient Second Order Neural Network Training Algorithms," *Lect. Notes Artif. Intell.*, vol. 3025, pp. 230–245, 2004.

[17] I. R. Horowitz, "Obstetrics and Gynecology Clinics of America: Preface," *Obstetrics and Gynecology Clinics of North America*, vol. 28, no. 4, 2001.

[18] E. J. Crosbie, M. H. Einstein, S. Franceschi, and H. C. Kitchener, "Human papillomavirus and cervical cancer.," *Lancet*, vol. 382, no. 9895, pp. 889–99, 2013.

[19] N. A. Mat-Isa, M. Y. Mashor, and N. H. Othman, "An automated cervical pre-cancerous diagnostic system," *Artif. Intell. Med.*, vol. 42, no. 1, pp. 1–11, 2008.

[20] C. M. Bishop, *Pattern Recognition and Machine Learning*, vol. 4, no. 4. 2006.

[21] Y. Song *et al.*, "Accurate cervical cell segmentation from overlapping clumps in pap smear images," *IEEE Trans. Med. Imaging*, vol. 36, no. 1, pp. 288–300, 2017.

[22] B. Ashok and P. Aruna, "Comparison of Feature selection methods for diagnosis of cervical cancer using SVM classifier," vol. 6, no. 1, pp. 94–99, 2016.

[23] H. Lee and J. Kim, "Segmentation of Overlapping Cervical Cells in Microscopic Images with Superpixel Partitioning and Cell-Wise Contour Refinement," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2016, pp. 1367–1373.

[24] J. Su, X. Xu, Y. He, and J. Song, "Automatic Detection of Cervical Cancer Cells by a Two-Level Cascade Classification System.," *Anal. Cell. Pathol. (Amst).*, vol. 2016, p. 9535027, 2016.

[25] M. Sharma, S. Kumar Singh, P. Agrawal, and V. Madaan, "Classification of Clinical Dataset of Cervical Cancer using KNN," *Indian J. Sci. Technol.*, vol. 9, no. 28, 2016.

[26] R. Kumar, R. Srivastava, and S. Srivastava, "Detection and Classification of Cancer from Microscopic Biopsy Images Using Clinically Significant and Biologically Interpretable Features," vol. 2015, 2015.

[27] Y. Song, L. Zhang, S. Chen, D. Ni, B. Lei, and T. Wang, "Accurate segmentation of cervical cytoplasm and nuclei based on multiscale convolutional network and graph partitioning," *IEEE Trans. Biomed. Eng.*, vol. 62, no. 10, pp. 2421–2433, 2015.

[28] T. Chankong, N. Theera-Umpon, and S. Auephanwiriyakul, "Automatic cervical cell segmentation and classification in Pap smears," *Comput. Methods Programs Biomed.*, vol. 113, no. 2, pp. 539–556, 2014.

[29] Y. Song *et al.*, "A deep learning based framework for accurate segmentation of cervical cytoplasm and nuclei," *Conf. Proc. ... Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. IEEE Eng. Med. Biol. Soc. Annu. Conf.*, vol. 2014, pp. 2903–2906, 2014.

[30] J. Talukdar, C. K. Nath, and P. H. Talukdar, "2013-

Fuzzy Clustering Based Image Segmentation of Pap smear Images of Cervical Cancer Cell Using FCM Algorithm.pdf," vol. 3, no. 1, pp. 460–462, 2013.

[31] Z. Lu, G. Carneiro, and A. P. Bradley, "An improved joint optimization of multiple level set functions for the segmentation of overlapping cervical cells," *IEEE Trans. Image Process.*, vol. 24, no. 4, pp. 1261–1272, 2015.

[32] Z. Lu, G. Carneiro, and A. P. Bradley, "Automated nucleus and cytoplasm segmentation of overlapping cervical cells," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 8149 LNCS, no. PART 1, pp. 452–460, 2013.

[33] A. Gençtav, S. Aksoy, and S. Önder, "Unsupervised segmentation and classification of cervical cell images," *Pattern Recognit.*, vol. 45, no. 12, pp. 4151–4168, 2012.

[34] A. Kale and S. Aksoy, "Segmentation of cervical cell images," in *Proceedings - International Conference on Pattern Recognition*, 2010, pp. 2399–2402.

[35] C. Bergmeir, M. García Silvente, and J. M. Benítez, "Segmentation of cervical cell nuclei in high-resolution microscopic images: A new algorithm and a web-based software framework," *Comput. Methods Programs Biomed.*, vol. 107, no. 3, pp. 497–512, 2012.

[36] P. Y. Pai, C. C. Chang, and Y. K. Chan, "Nucleus and cytoplast contour detector from a cervical smear image," *Expert Syst. Appl.*, vol. 39, no. 1, pp. 154–161, 2012.

[37] I. Muhimmah, R. Kurniawan, and Indrayanti, "Automated cervical cell nuclei segmentation using morphological operation and watershed transformation," in *Proceeding - 2012 IEEE International Conference on Computational Intelligence and Cybernetics, CyberneticsCom 2012*, 2012, pp. 163–167.

[38] K. Li, Z. Lu, W. Liu, and J. Yin, "Cytoplasm and nucleus segmentation in cervical smear images using Radiating GVF Snake," *Pattern Recognit.*, vol. 45, no. 4, pp. 1255–1264, 2012.

[39] I. Journal, C. Applications, and T. Bangalore-, "Papsmear Image based Detection of Cervical Cancer," vol. 45, no. 20, pp. 35–40, 2012.

[40] H. Kong, M. Gurcan, and K. Belkacem-Boussaid, "Partitioning histopathological images: An integrated framework for supervised color-texture segmentation and cell splitting," *IEEE Trans. Med. Imaging*, vol. 30, no. 9, pp. 1661–1677, 2011.

[41] L. Zhang, S. Chen, T. Wang, Y. Chen, S. Liu, and M. Li, "A Practical Segmentation Method for Automated Screening of Cervical Cytology," *2011 Int. Conf. Intell. Comput. Bio-Medical Instrum.*, no. December, pp. 140–143, 2011.

[42] M. E. Plissiti, C. Nikou, and A. Charchanti, "Combining shape, texture and intensity features for cell nuclei extraction in Pap smear images," *Pattern Recognit. Lett.*, vol. 32, no. 6, pp. 838–853, 2011.

[43] S. N. Sulaiman, N. A. M. Isa, I. A. Yusoff, and N. H. Othman, "Overlapping cells separation method for cervical cell images," in *Proceedings of the 2010 10th International Conference on Intelligent Systems Design and Applications, ISDA'10*, 2010, pp. 1218–1222.

[44] C. H. Lin, Y. K. Chan, and C. C. Chen, "Detection and segmentation of cervical cell cytoplast and nucleus," *Int. J. Imaging Syst. Technol.*, vol. 19, no. 3, pp. 260–270, 2009.

[45] Y. Marinakis, G. Dounias, and J. Jantzen, "Pap smear diagnosis using a hybrid intelligent scheme focusing on genetic algorithm based feature selection and nearest neighbor classification," *Comput. Biol. Med.*, vol. 39, no. 1, pp. 69–78, 2009.

[46] M. E. Plissiti, E. E. Tripoliti, A. Charchanti, O. Krikoni, and D. I. Fotiadis, "Automated detection of cell nuclei in pap stained cervical smear images using fuzzy clustering," in *IFMBE Proceedings*, 2008, vol. 22, pp. 637–641.

[47] S. F. Yang-Mao, Y. K. Chan, and Y. P. Chu, "Edge enhancement nucleus and cytoplast contour detector of cervical smear images," *IEEE Trans. Syst. Man, Cybern. Part B Cybern.*, vol. 38, no. 2, pp. 353–366, 2008.

[48] V. Pomero, D. Mitton, S. Laporte, J. A. De Guise, and W. Skalli, "Fast accurate stereoradiographic 3D-reconstruction of the spine using a combined geometric and statistic model," *Clin. Biomech.*, vol. 19, no. 3, pp. 240–247, 2004.

[49] J. Y. Zhang and Y. X. Liu, "Cervical cancer detection using SVM based feature screening," in *MEDICAL IMAGE COMPUTING AND COMPUTER-ASSISTED INTERVENTION - MICCAI 2004, PT 2, PROCEEDINGS*, 2004, vol. 3217, no. 2, pp. 873–880.

[50] F. Ritter *et al.*, "Medical image analysis," *IEEE Pulse*, vol. 2, no. 6, pp. 60–70, 2011.