Cervical Cancer: Open Access

Sarwar, Cervical Cancer 2016, 1:1
DOI: 10.4172/2475-3173.1000105

Research Article Open Access

Artificial Intelligence Based Semi-automated Screening of Cervical Cancer Using a Primary Training Database

Abid Sarwar*

Department of Computer Science and IT, University of Jammu, Jammu, India

Abstract

Objective: The primary objective of this research work is to develop a novel benchmark database of digitized and calibrated, cervical cells obtained from slides of Papanicolaou smear test, which is done for screening of cervical cancer. This database can serve as a potential tool for designing, developing, training, testing and validating various Artificial intelligence based systems for prognosis of cervical cancer by characterization and classification of Papanicolaou smear images. The database can also be used by other researchers for comparative analysis of working efficiencies of various machine learning and image processing algorithms. The database can be obtained by sending a request to the corresponding author. Besides developing a rich machine learning database we have also presented a novel artificial intelligence based hybrid ensemble technique for efficient screening of cervical cancer by automated analysis of Papanicolaou smear images.

Methodology: The correct and timely diagnosis of cervical cancer is one of the major problems in the medical world. From the literature it has been found that different pattern recognition techniques can help them to improve in this domain. Papanicolaou smear (also referred to as Pap smear) is a microscopic examination of samples of human cells scraped from the lower, narrow part of the uterus, called cervix. A sample of cells after being stained by using Papanicolaou method is analyzed under microscope for the presence of any unusual developments indicating any precancerous and potentially precancerous developments. Abnormal findings, if observed are subjected to further precise diagnostic subroutines. 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. It is the most popular technique used for screening of cervical cancer. Pap smear test, if done with a regular screening programs and proper follow-up, can reduce cervical cancer mortality by up to 80% . The contribution of this paper is that we have created a rich machine learning database of quantitatively profiled and calibrated cervical cells obtained from Pap-smear test slides. The database so created consists of data of about 200 clinical cases (8091 cervical cells), which have been obtained from multiple health care centers, so as to ensure diversity in data. The slides were processed using a multi-headed digital microscope and images of cervical cells were obtained, which were passed through various data preprocessing subroutines. After preprocessing the cells were morphologically profiled and scaled to obtain separate quantitative measurements of various features of cytoplasm and nucleus respectively. The cells in the database were carefully classified in different corresponding classes according to latest 2001-Bethesda system of classification, by technicians. In addition to this, we have also pioneered to apply a novel hybrid ensemble system to this database in order to evaluate the effectiveness of both novel database and novel hybrid ensemble technique to screen cervical cancer by categorization of Pap smear data. The paper also presents a comparative analysis of multiple artificial intelligence based classification algorithms for prognosis of cervical cancer.

Results: For evaluating the effectiveness and correctness of the digital database prepared in this work, authors implemented this database for training, testing and validating fifteen different artificial intelligence based machine learning algorithms. All algorithms trained with this database presented commendable efficiency in screening of cervical cancer. For two-class problem all the algorithms trained with the digital database showed the efficiencies in range of about 93-95% while as in case of multi class problem algorithms expressed the efficiencies in the range of about 69-78%. The results indicate that the novel digital database prepared in this work can be efficiently used for developing new machine learning based techniques for automated screening of cervical cancer. The results also indicate that hybrid ensemble technique is an efficient method for classification of pap-smear images and hence can be effectively used for diagnosis of cervical cancer. Among all the algorithms implemented, the hybrid ensemble approach outperformed and expressed an efficiency of about 98% for 2-class problem and about 86% for 7-class problem. The results when compared with the all the standalone classifiers were significantly better for both two-class and multi-class problems.

Keywords: Artificial intelligence; Cervical cancer; Medical database; Ensemble technique; Pap smear test

Introduction

Timely and precise diagnosis of cervical cancer is an important real-world medical problem. Cervical cancer has turn into one of the main causes of mortality among women around the globe and it has become a major concern among the scientific community to investigate into it, leading to early diagnosis and mitigated mortality rate. Cervical cancer is the fourth most common cancer in women, and the seventh overall in the world. As reported by WHO the toll of cervical cancer in 2012 was

***Corresponding author:** Abid Sarwar, Department of Computer Science and IT, University of Jammu, Jammu, India, Tel: 91 9697436894; E-mail: sarwar.aabid@gmail.com

Received December 17, 2015; **Accepted** December 29, 2015; **Published** January 05, 2016

Citation: Sarwar A (2016) Artificial Intelligence Based Semi-automated Screening of Cervical Cancer Using a Primary Training Database. Cervical Cancer 1: 105. doi: 10.4172/2475-3173.1000105

528000 with the number of deaths equal to 266000. A large majority of these cases (around 84%) are reported from the developing and under-developed countries as compared to the developed countries, which is attributed to comparatively poor access to screening and treatment services. Over the last few decades Artificial intelligent techniques have been increasingly used in solving problems in medical domains such as in Oncology [1-5] , Urology [4], Liver Pathology [6], Cardiology [7,8], Gynecology [9], Thyroid disorders [10,11], Perinatology [12] etc. The diagnosis of a large segment of diseases is based upon the pathological tests done by the patient. Owing to this reason, use of AI based classifier systems is gaining immense importance now days. The primary purpose of employing AI in medicine is creation of such artificially intelligent systems which can provide assistance to a medical doctor in delivering expert diagnosis. These artificial intelligent systems support the clinical decision making by anticipating the diagnostic results, after being trained using previously acquired training data followed by expending specific information of some patient case. The use of Artificial intelligence in medicine has shown substantial progress in achieving timely, reliable diagnosis and more precise treatment of many diseases. The cervical cancer etiology is still not clear and medical experts have not come up with a single dominant cause. Prevention is not so unproblematic and early detection is the only means to help the patients to survive. If the cancerous cells are detected before spreading to other organs, the survival rate for patients is more than 97% (American Cancer Society Hompage, 2008).

Artificial intelligence

Artificial intelligence (AI) is a subpart of computer science, concerned with, how to give computers the sophistication to act intelligently, and to do so in increasingly wider realms. The field was founded on the claim that a central property of humans, Intelligence can be simulated by a machine. Artificial Intelligence has now days become an essential part of the technology industry, providing the heavy lifting for many of the most difficult problems from all walks of daily life. These days, many efforts are being laid upon the development of models of diseases, using the synthetic intelligence to overcome the difficulties faced using the traditional rule based modeling techniques. Such intelligent models of diseases have resulted in significant progress in our understanding of the progression of various disorders, and thus helped in gaining more clinical expertise. Efforts to develop such programs have led to substantial progress in our understanding of clinical expertise, in the translation of such expertise into cognitive models, and in the conversion of various models into promising experimental programs. Of equal importance, these programs have been steadily improved through the correction of flaws shown by confronting them with various clinical problems.

Cervical cancer

Cervical cancer is malignant tumor that occurs when the cervix tissue cells begin to grow and replicate abnormally without controlled cell division and cell death. In such a state, the body is unable to use and manage such cells for carrying out usual function as a result of which these cells transform into a tumor. If the tumor is malignant, its cell flow through the blood stream and spread to other parts of body, as a result those parts also get infected. Usually the cervical cancer takes number of years to develop. These infected cells are then distinguished as cervical intra-epithelial neoplasia (CIN) or cervical dysplasia. The cells over the surface of cervix that show unusual changes and potentially precancerous developments are called CIN. In most of the cases CIN remains stable, or these are eliminated by host's immune system response. Although, a small percentage of cases progress to become cervical cancer, if not treated. Studies have found that CIN usually results from a virus called human papillomavirus (HPV) which is generally sexually transmitted. Although there are more than 120 types of HPV known [13], only a 15 are classified as high-risk types (16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, 59, 68, 73, and 82) [14], 3 as probable-high-risk (26, 53, and 66), and 12 as low-risk (6, 11, 40, 42, 43, 44, 54, 61, 70, 72, 81, and CP6108).In many cases even after getting infected with HPV, it is generally eliminated by the response of host's immune system, but in many cases many cases, where HPV is not done away with by the immune system of host, it may develop into cervical cancer. The common risk factors linked with cervical cancer include first intercourse at an early age, pregnancy at early age, having sex with multiple partners, weak immune system, smoking , use of oral contraceptives, improper menstrual hygiene etc. At an early stage the cervical cancer may be completely asymptomatic. The early stages of the cervical cancer are usually asymptomatic but symptoms do appear with the progression of pre-cancer to invasive cancer and typically shows vaginal bleeding abnormally, vaginal discharge, pain during vaginal intercourse. New bleeding may be experienced by the women who have had their menopause. Cervical cancer is the second most commonly diagnosed [15] and fifth deadliest [16] cancer in women throughout the world. In developing countries cervical cancer has a major share in cancer mortality [17]. Every year about 5,00,000 new cases are diagnosed and among which about 2 50 000 patient die. Because of poor access to screening and treatment services, approximately 80% of this disease occurs in women living in low- and middle-income countries [18]. Because of poor access to screening and treatment services, approximately 80% of this disease occurs in women living in low- and middle-income countries [18].

Papanicolaou test

The Papanicolaou test (Pap smear) has been the widely used method in cervical cancer screening for many decades and has showed a dramatic lowering of incidents of cervical cancer and hence in related mortality rates in many countries [17]. In taking a Pap smear, the cells of are scraped form the outer opening of the cervix for microscopic examination and to lookup for irregularities. The aim of the test is to detect any pre-cancerous or potentially pre-cancerous alterations called cervical intraepithelial neoplasia (CIN) or cervical dysplasia. Pap test is also used to detect endocervix and endometrium abnormalities and infections. In many developed countries, regular Pap smear screening is highly recommended for females who have had frequent sex with multiple partners [19]. If any unusual findings are observed the test may need to be repeated within a year. If the abnormality observed requires closer examination, a detailed cervical inspection by colposcopy may be done. HPV DNA testing may also be suggested to such patients, which acts as a supplementary to Pap smear testing. Once the sample is obtained, Papanicolaou technique is used to stain it. Staining using this technique helps to differentiate the cells in smear preparation from various other bodily secretions as unstained cells cannot be seen under a simple compound microscope. Most of the abnormal results are mildly abnormal (called low-grade squamous intraepithelial lesion (LSIL)) which indicates HPV infection. Most low-grade cervical dysplasia relapse by their own without usually causing cervical cancer, but presence of dysplasia can act as a warning that greater monitoring is needed. Generally, some of Pap results are high-grade squamous intraepithelial lesion (HSIL), and very few of them indicate cancer.

Bethesda system

The Bethesda system is a standard system, used worldwide for

reporting diagnosis of cervicovaginal cytology. It was first introduced in 1988 with an aim to standardize the terminology used and establish more consistent reports for reporting, which would facilitate clear guidelines for clinical management. Since its first publication in 1988, it was revised two times, in April 1991 and April 2001. The 2001-Bethesda system consists of several components and subcomponents, as summarized in (Table 1).

Literature Review

Artificial intelligence based algorithms are increasingly being used to analyze and interpret large volumes of data for solving problems in medical domains [20], as is evident from a considerable amount of research done in this field during recent past [21] have proposed a model (Levenberg–Marquardt feedforward MLP neural network) for classification of cervical cell images. This model is a novel model based upon the feed forward neural network, and has been trained and tested from the data obtained from 100 patients. The model is composed of two stages, in the first stage, images are preprocessed to reduce the noises (if any) without compromising on the resolution of images and in the second stage image processing algorithms are applied to cell images to achieve a linear plot, which were then used as LMFFNN inputs for classification of cervical cell images. Their model has shown 100% correct classification rate and was thus found to be in good correlation with the decision made by the medical experts [22] developed an automated system for diagnosis and screening pre-cancerous cervical cells. Their system was composed of two components one of the components for automatic feature extraction and the other for an intelligent diagnosis. The first component automatically draws out four critical features (i.e. nucleus size, cytoplasm size, nucleus grey level and cytoplasm grey level) from the cervical cells. The authors have developed a novel algorithm called region-growing-based features extraction (RGBFE). This algorithm is used for extraction of features important for diagnosis. The data about all these features when extracted from the cervical cell images are fed to the intelligent diagnostic part. The pre-cancerous stages are predicted using artificial neural network developed using a novel architecture called hybrid multilayered perceptron (H2MLP) network. The cells are classified into three classes as normal, low grade intra-epithelial squamous lesion (LSIL) and high grade intra-epithelial squamous lesion (HSIL).The capability of the system so developed is assed using 550 clinical cases which were classified as normal cases to be 211, LSIL cases to be 143 and HSIL cases to be 196. For the purpose of evaluation of performance of the system in comparison to the manual extraction by expert cytologist, correlation test was used. The results imply a strong linear relationship between mean of grey level and the estimated size with that extracted the expert cytotechnologist [23] developed an intelligent system for automatic detection of shape of nucleus and cytoplasm of cell of cervix obtained from Pap smear. They called this system as nucleus and cytoplasm contour detector (NCC detector). They used adaptable threshold decision method to distinguish the cell from the cervical smear image, followed by using the maximal gray-level-gradient-difference method, proposed by them, for extraction of the nucleus from the cell. The comparative analysis of NCC detector with the earlier available methods reveled that NCC detector performs better than the edge enhancement nucleus and cytoplast contour detector model and the then gradient vector flow-active contour model [24] proposed an unsupervised approach for segmentation and classification of cervical cells obtained from Pap smear slides. The segmentation process involves providing an automatic threshold for separating the cell regions from the background, a multi-scale hierarchical segmentation algorithm to partition these regions based on homogeneity and circularity, and a binary classifier to finalize the separation of nuclei from cytoplasm within the cell regions. The proposed procedure constructs a tree using hierarchical clustering, and then arranges the cells in a linear order by using an optimal leaf ordering algorithm that maximizes

Reporting	Criteria
Specimen Adequate/ Inadequate	Specimen is considered adequate for evaluation on the basis of presence or absence of endocervical or transformation zone component and other quality indicators e.g., partially obscuring blood, inflammation, etc.
Interpretation of Observations	
a. Negative for Intraepithelial lesion or malignancy	This is reported when there is no cellular evidence of neoplasia. It is also specified-whether or not there are organisms found or any other non-neoplastic findings.
➢ Organisms	• *Trichomonas vaginalis* • Shift in flora suggestive of bacterial vaginosis • Bacteria morphologically consistent with *Actinomyces* spp. • Fungal organisms morphologically consistent with *Candida* spp. • Cellular changes consistent with *Herpes simplex* virus
➢ Other Non-neoplastic findings	• Reactive cellular changes associated with: - Inflammation (includes typical repair) - Radiation - Intrauterine device (IUD) • Glandular cells status posthysterectomy • Atrophy
➢ Others	• Endometrial cells (in a woman ≥40 years of age)
b. Epithelial Cell Abnormalities	Squamous Cell: • Atypical squamous cells - of undetermined significance (ASC-US) - cannot exclude HSIL (ASC-H) • Low grade Squamous intra epithelial lesion (LSIL) (encompassing: HPV/ mild Displasia/ CIN 1) • High grade Squamous intra epithelial lesion (LSIL) (encompassing: moderate and severe Displasia, CIS, CIN 2 and 3) • Squamous cell carcinoma Glandular Cell: • Atypical glandular cells (AGC) (endocervical cells, endometrial cells or not otherwise specified) • Atypical glandular cell, favor neoplastic (endocervical cells or not otherwise specified) Endocervical adenocarcinoma *in situ* (AIS) Adenocarcinoma
c. Others	Endometrial cells in a woman ≥40 years of age

Table 1: Summary of various component and sub components of 2001-Bethseda system of classification.

the similarity of adjacent leaves without any requirement for training examples or parameter adjustment. Performance evaluation using two data sets show the effectiveness of the proposed approach in images having inconsistent staining, poor contrast, and overlapping cells [25] have performed a case study and have prepared data and baseline for comparing classification methods. The data collected by them consists of 917 images of Pap-smear cells, classified carefully by cytotechnicians and doctors. Each cell is described by 20 numerical features, and the cells fall into 7 classes. Authors have also done a basic data analysis that includes scatter plots and linear classification results, in order to provide domain knowledge and lower bounds on the acceptable performance of future classifiers.

Database of digitized and calibrated Pap-Smear cells

For any machine learning algorithm, the database with which it is trained plays an important role. It is said that a machine can be made to learn and reproduce any human behavior, provided it is trained with suitably precise database. The database prepared in this work consists of 8091 tuples which represent data of about 200 clinical cases; each tuple containing 40 attributes, and is identified by a unique primary key. Among the 41 attributes 19 correspond to the features of Cytoplasm, 19 represent the features of Nucleus, one attribute about the ratio of Nuclear-Cytoplasmic area and the one last attribute identifies the class to which the particular cell belongs. Each tuple corresponding to one cervical cell, extracted from the slides of pap-smear test which were obtained from three medical health care institutions in northern India viz Government Medical College, Jammu, Acharaya Shree Chandra College of Medical Science and Hospital, Jammu and Nijjer Pathology Laboratory, Amritsar. Form the hospital records; cases of cervical cancer were identified and their corresponding pathological records were obtained. All the medical ethical issues were taken in to consideration so that the identity of the patient is not compromised and reveled in any case. The slides were observed and analyzed under a multi-headed microscope (NIKON Nikon Eclipse E400 DS-F12) having a digital camera mounted over it and connected and configured with an attached computer. After examining the cells under the microscope under different level of magnifications (i.e. 10x, 40x, 100x), images of all the slides were captured at 40x magnification, so as to ensure uniformity and consistency among all the cells.

These images were allowed to pass from various pre-processing subroutines so as to obtain distinct individual cells segregated from the cell clusters. While preprocessing the images care was taken that the size and resolution of the images is preserved. The individual cells were cropped-off from the cell cluster obtained from the microscope followed by enhancing their brightness, color and contrast using image processing techniques, where ever required, so as to make the Cytoplasmic and Nuclear features easily recognizable. Unique names were assigned to all the cells with an aim to identify each of them distinctly in the database. The database as such, contains a total of 7,000 cervical cells which have been carefully differentiated manually into different classes respectively using the 2001-Bethesda system of classification. To ensure the accuracy each cell included in the database was inspected by two trained Cyto-pathologists, and complex samples were also subjected to multiple subroutines of examination. The diagnosis done by the cyto-pathologists were cross checked with the corresponding diagnosis in the medical records for the clinical case. In case of any difference of opinion in Pap smear reporting the sample was excluded from database. As such the final database so prepared contains diagnoses which are most precise, certain and accurate. After discussion with the medical experts of the concerned domain, authors identified 39 morphological features of cervical cells, 19 each from the Cytoplasm and Nucleus, on the basis of which the cells were profiled. For extraction of these features an open source software, CellProfiler was used. It was developed by BROAD institute and Massachusetts Institute of Technology, to enable biologists to quantitatively measure the phenotypes of medical images. The software consists of a number of discrete individual modules which can be arranged sequentially to make a tailored pipeline of activities as desired by the analysis. This pipeline of activities thus made, is used to look-up and quantitatively measures the features of biological objects of interest in any input image. After processing all the cells on the cell profiling utility, the data obtained was arranged in a spreadsheet and all the properties of database were incorporated, so that digital data could be easily correlated with the corresponding clinical case. For evaluating the effectiveness and correctness of the database so prepared, authors implemented this database for training, testing and validating fifteen different artificial intelligence based machine learning algorithms. All the algorithms trained with this database presented a commendable efficiency in screening of cervical cancer (Figure 1). show the block diagram of how the clinical data was transformed into a digital database. Table 2 summarizes the number of different types of Pap smear images of each class in the database, Figure 2 shows sample images from the database and Figure 3 shows the screen short of the digital database. Among these seven classes the first four i.e. 1, 2, 3 and 4 are classified as normal cells whereas the last three i.e. 5, 6 and 7 are classified as abnormal cells. The screening of these Pap smear images can be regarded as classification into normal and abnormal cells i.e. two-class problem and also as an elaborated classification into seven respective classes.

Features selected for calibration of cervical cells and their prognostic significance

Automated cervical screening system for the detection of malignancy or any other potentially pre-malignant abnormalities requires a detailed analysis of morphological parameters of the cells of cervix. The set of attributes and features chosen for the quantitative evaluation have to be such, as they should be measurable in specific time with good accuracy, re-measurable and reflect the properties of the cell in a very precise manner [26]. Morphological examination of a cell includes the evaluation of cell in terms of the properties of its size, shape, orientation, proportion of area of cell shared by nucleus and cytoplasm etc. The features selected for evaluation in this study cover both the shape and size aspects of morphological analysis of the cell. Among the 39 selected features, 15 correspond to the measurement of the size and dimension whereas the other 24 correspond to measurement of shape and contour of the cytoplasm and nucleus of the cell. Computer techniques based on various image processing algorithms provide simple and proficient formulization of these morphological features as compared to the manual interpretation by cyto-pathologists which are subjective, laborious and many times prone to human errors [27,28]. In addition to this many different methods are available for the separation of nucleus and cytoplasm from the image of a cell; as such nuclear and Cytoplasmic features can be independently analyzed in detail. The Cytoplasmic and Nuclear parameters which were selected are as listed in the Table 3. The calibration of these features was done in terms of number of pixels.

Among the parameters, Area of Cytoplasm/Nucleus defines the total number of pixels enclosed in the area occupied by the Cytoplasm/Nucleus. The Area of cytoplasm and Area of nucleus contributes a lot in prognosis of any malignance related disorder. It has been observed that in normal cervical cells cytoplasm has a larger area and nucleus

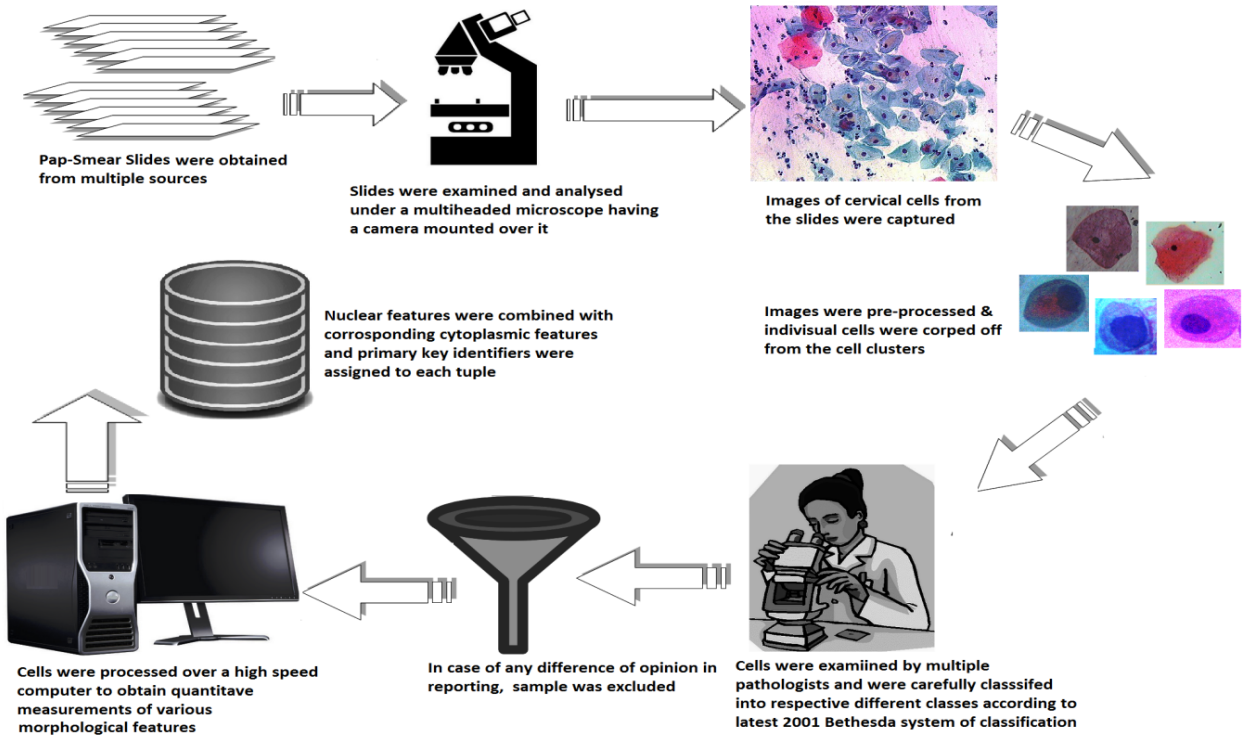

Figure 1: Block diagram of various steps showing how the clinical data from the Pap smear slides was transformed into digital database.

Class	Category	Type of cell	Number of Cell images	Sub total
1	Normal	Superficial squamous	1,474	4,956
2		Intermediate squamous	1,354	
3		Parabasal squamous	1,056	
4		Basal squamous	1,072	
5	Abnormal	Low-grade squamous intraepithelial lesion	1,002	3,135
6		High-grade squamous intraepithelial lesion	1,122	
7		Squamous cell carcinoma	1,011	

Table 2: Summary of images of each type of cells.

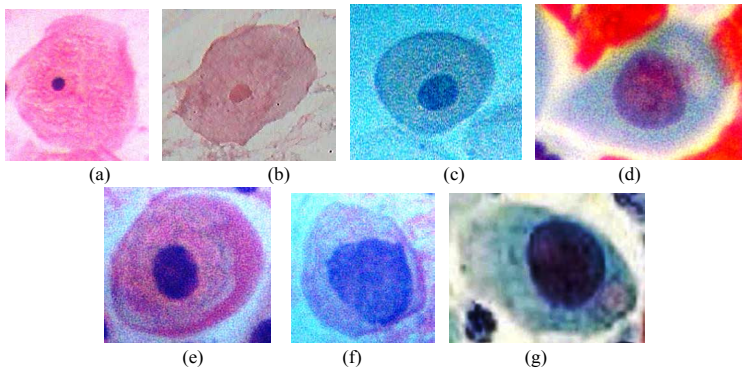

Figure 2: (a) Superficial squamous cells (b) Intermediate squamous cells (c) Parabasal squamous cell (d) Basal squamous cell (e) Low-grade squamous intraepithelial lesion (f) High-grade squamous intraepithelial lesion (g) Squamous cell carcinoma.

has a relatively lesser area, in contrast to the abnormal cervical cells (CIN-II and CIN-III) which have large nucleus area almost equal to that of cytoplasm. Meyer-Arendt and Humphreys (1972) observed that cancerous cells are smaller in area than normal superficial and intermediate cells. Cytoplasmic and nuclear perimeter defines the number of pixels that are covered along the boundary of the region of cytoplasm and nucleus. The perimeter has a contributing role as the normal cervical cells have larger Cytoplasmic Perimeter value and low

Citation: Sarwar A (2016) Artificial Intelligence Based Semi-automated Screening of Cervical Cancer Using a Primary Training Database. Cervical Cancer 1: 105. doi: 10.4172/2475-3173.1000105

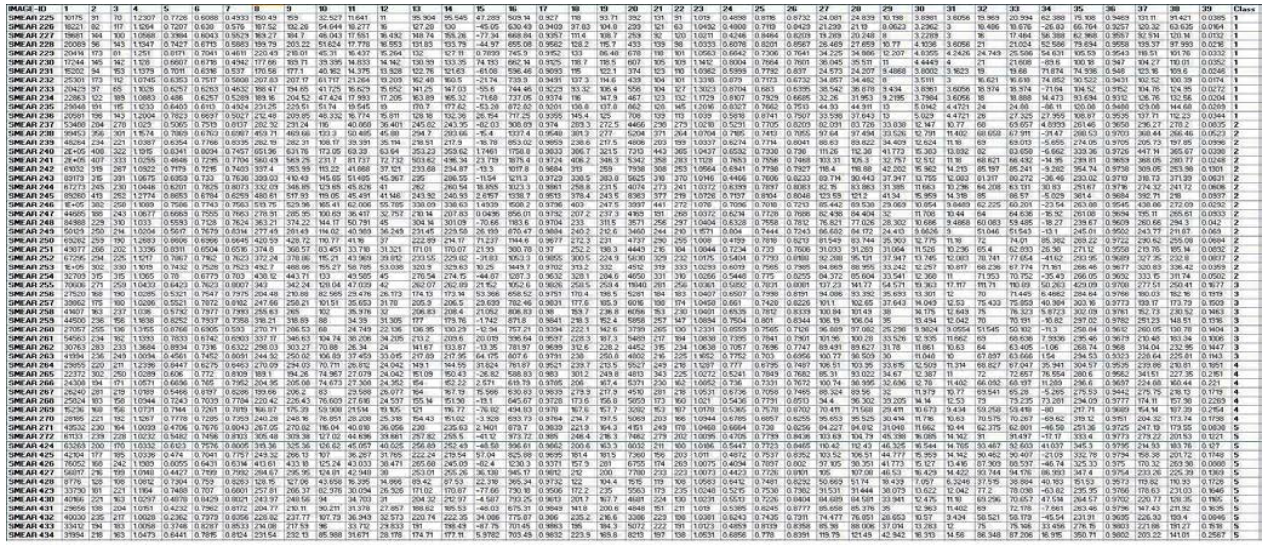

Figure 3: Screen-short of a sample of database showing the Image-id, various attributes and their corresponding classes.

Features of Cytoplasm | Title in database	Features of Nucleus | Title in database
Area of the Cytoplasm | CArea	Area of the Nucleus | NArea
Center of Cytoplasm X-axis | CCenter_X	Center of Nucleus X-axis | NCenter_X
Center of Cytoplasm Y-axis | CCenter_Y	Center of Nucleus Y-axis | NCenter_Y
Compactness of Cytoplasm | CCompactness	Compactness of Nucleus | NCompactness
Eccentricity of Cytoplasm | CEcentricity	Eccentricity of Nucleus | NEcentricity
Extent of Cytoplasm | CExtent	Extent of Nucleus | NExtent
Form factor of Cytoplasm | CFormFactor	Form factor of Nucleus | NFormFactor
Major Axis Length of Cytoplasm | CMajorAxisLength	Major Axis Length of Nucleus | NMajorAxisLength
Maximum Radius of Cytoplasm | CMaximumRadius	Maximum Radius of Nucleus | NMaximumRadius
Mean Radius of Cytoplasm | CMeanRadius	Mean Radius of Nucleus | NMeanRadius
Median Radius of Cytoplasm | CMedianRadius	Median Radius of Nucleus | NMedianRadius
Minor Axis length of Cytoplasm | CMinorAxisLength	Minor Axis length of Nucleus | NMinorAxisLength
Cytoplasmic Orientation | COreantation	Nuclear Orientation | NOreantation
Cytoplasmic perimeter | CPerimeter	Nuclear perimeter | NPerimeter
Solidity of Cytoplasm | CSolidity	Solidity of Nucleus | NSolidity
Location of Center of Cytoplasm X-Axis | CLocation_Center_X	Location of Center of Nucleus X-Axis | NLocation_Center_X
Location of Center of Cytoplasm X-Axis | CLocation_Center_Y	Location of Center of Nucleus X-Axis | NLocation_Center_Y
Maximum Farthest diameter of Cytoplasm | CMaxFeretDiameter	Maximum Farthest diameter of Nucleus | NMaxFeretDiameter
Minimum Farthest diameter of Cytoplasm | CMinFeretDiameter	Minimum Farthest diameter of Nucleus | NMinFeretDiameter
Nucleus-Cytoplasm Ratio | NCRatio	

Table 3: List of parameters along with their representation in database, which were used to profile the cervical cells.

Nuclear Perimeter value whereas the cells affected by cervical inter-epithelial neoplasia both large values for the perimeter of cytoplasm and nucleus. Form Factor represents the roundness of an object and is calculated using the formulae: $4*\pi*Area/Perimeter^2$. The value of form factor is equal to 1, for objects which are perfectly circular in shape. Form factor measures the deviation of the shape of the cell from the normal circular to distorted irregular, as such can be used to identify the uneven and asymmetric cells. The Solidity is measured as proportion of the area under the convex hull to Area of the object. The convex hull is the smallest convex polygon containing all points in the object. The value of Solidity is equal to 1 for smooth and regular surfaces of Normal and Atypical Squamous Cells of undetermined significance (ACS-US) category cells whereas cells of High grade squamous intraepithelial lesion (HSIL) encompassing moderate (CIN-II) and severe dysplasia (CIN-III) category cells have very low solidity value owing to their, irregular and convex surfaces. Cytoplasmic and Nuclear Extent

symbolizes the area of the whole of the figure which is covered by the object, and is calculated by dividing the Cytoplasmic/Nuclear area by the area of the figure. Center of Cytoplasm/Nucleus X-Axis and Y-Axis gives the X and Y coordinates of the point which is farthest away from the edge of the cytoplasm/nucleus. It has the significance that it represents the morphological shape of the cytoplasm and nucleus. It also measures the relative position of nucleus with respect to the center of the cytoplasm. The parameter Eccentricity is the measure of roundness or oblongivity of cytoplasm and nucleus and is measured by ratio of Foci of the ellipse to the major axis length. With the progression of Cervical intra epithelial neoplasia, the nucleus becomes distorted with the result of which the value of eccentricity also increases and approaches to 1 in a very oblong nucleus. Eccentricity thus provides a good measure, about the irregular cellular shapes of High grade squamous intraepithelial lesion (HSIL) encompassing moderate (CIN-II) and severe dysplasia (CIN-III) category. Major Axis and Minor Axis Length of Cytoplasm/Nucleus signifies in pixel length of the major axis and minor axis respectively, of the ellipse that best fits the object of interest. This measure also helps in computing the other features like elongation and eccentricity. Cytoplasmic and Nuclear Orientation specifies the angle between the X-axis and the major axis of the ellipse that fits best the biological object of interest.Cytoplasm/Nuclear Compactness is used for calibration of shape and this parameter is widely utilized to describe the morphometric changes in the biological entities [29,30]. It is measured by calculating the ratio of square of perimeter to area of the ellipse. It has been observed that the cells of the ecto-cervix (mature superficial cells and intermediate cells) have more compact nucleus but less compact cytoplasm as compared to immature squamous cells (Para basal cells and basal cells) which are characterized by both compact cytoplasm and compact nucleus. It has also been observed that the presence of Malignancy can be detected by the value of Nucleus:Cytoplasm ratio, as the abnormal cells (Atypical Squamous Cells cannot exclude HSIL, CIN-II and CIN-III) have high values of Nucleus:Cytoplasm ratio and may approach to 1:1,asthe Normal cells (Normal, Atypical Squamous Cells of undetermined significance) have relatively lower values of Nucleus:Cytoplasm ratio which may approach to 1:4 and 1:6. Similarly the values of MaxFeretDiameter, MinFeretDiameter, Maximum radius, Mean radius and Median radius have larger values for Normal polygonal shaped cervical cells and smaller for cells of dysplastic category (LSIL, CIN-II and CIN-III). Therefore the complete set of 39 above listed features of cytoplasm and nucleus together represent all the morphological features of prime importance for complete profiling of the individual cervical cell and thus contribute in precise and accurate automated screening of cervical cancer.

Novel hybrid ensemble technique

The classification by ensemble technique works by preparing a large set of discrete classifiers separately at the training time followed by considering their individual votes for framing the final output classification [31,32]. The method tends to boost the final predictive performance by coalescing the classifying potential of multiple learning algorithms which is better as compared to any of the individual constituent classifiers. Ensemble method not only boosts weak learners but also provides greater confidence to the classification process by considerably reducing the likelihood of misclassification of a particular instance by some algorithm. If an instance is wrongly classified by some classifier, the error is corrected by the right classification done by other contributing algorithms. Thus auto correcting of errors is achieved by nullifying the chance of wrong final classification by a classifier. The final output classification is selected by taking the votes for a particular instance from all individual prediction models taken under consideration. The hybrid ensemble system considered in this work is developed using fifteen different classification algorithms. The algorithms considered were Bagging, Decorate, Decision Table, Ensemble of Nested dichotomies (END), Filtered Classifier, J48 graft, Projective Adaptive Resonance Theory (PART), Multiple back propagation artificial neural network, Multiclass classifier, Naïve Bayes, Random subset space, Radial basis function network, Rotation Forest, Random forest and Random Committee. These algorithms were considered owing to their noted good performance for classification of complex datasets. The system can be viewed as an ensemble of ensemble classifiers as internally bagging, random forest, rotation forest, random committee, random subspace, decorate and nested dichotomies themselves are based on ensemble principle. Bagging stands for bootstrap aggregating, it divides the training data set into multiple subsets which may be overlapping, these subsets are further used to train multiple models and the final output is calculated by taking average or votes from the individual models. Random forest works by constricting many decision trees and for each input vector, a path in each classification tree is traversed for reaching a classification output. The same is recursively repeated for all instances. The final decision class for an instance is obtained by considering the maximum voting by all the trees in the forest so constructed. Random subspace is a generalization of random forest algorithm with a deviation that unlike random forest where decision trees alone are used as base classifiers; any type of classifier can be employed as individual classifier for building the ensemble. The system of nested dichotomies recursively converts a multi class classification problem into a simple binary classification problem. This binary classification problem is represented using binary trees. The final classification decision is obtained by aggregating the output of all possible candidate trees for the multi class problem. Random committee builds a diverse ensemble of multiple discrete tree classifiers each of which are trained using the same dataset but using a dissimilar seeds for generating randomness. The concluding predictions are then derived by taking mean probability approximations done by all the individual classifying trees. The DECORATE algorithm iteratively adds a learning classifier to the current ensemble at each iteration. The training data includes original data and some artificially constructed data. A new classifier trained on the new dataset is added to the ensemble and the final training error of the ensemble is noted for to accept or reject the new classifier added. The final classification for a test instance is done by consolidating the probability chances of an instance for to belong a particular class from all the classifiers involved in the ensemble. Rotation forest aims for an ensemble of accurate and diverse classifiers. Each of the classifiers is trained using different dataset and which are obtained by extraction of new feature subsets from the training data set using principle component analysis. Good diversity among the classifiers is achieved by different splits of the feature sets which considerably enhances the classifying capability of the ensemble.J48-Graft algorithm works by creating a grafted decision tree build previously using J48 tree algorithm. The grafting technique is an inductive process that appends new nodes replacing single leaf nodes or by grafting them in between leaves in inferred decision trees with the purpose of improving prediction accuracy. By grafting new nodes the tree finds alternate classifications for those regions which are either misclassified or not occupied by the training instances. Naïve bayes classifier works on the principle of conditional independence assumption, which means that it assumes that the presence or absence of some parameters of a class to be independent, to the presence or absence of some other parameters. Individual probabilities for all of the parameters for all the classes are calculated and final result

is consolidated to by multiplying their individual contributions. Test instance gets classified into the class having highest value for consolidated probability. Multiclass classifier solves a multiclass problem by transforming it into a simple two class problem which can be solved using a binary classifier. The binary classification problem is solved by using one-verses-all approach where n binary classifiers are designed to identify n classes. Among these n classifiers the classifier ***i*** is trained to identify the tuples belonging to class ***i*** as positive and rest as negative. Same is repeated for other classifiers. Final results are consolidated by voting from all the *n* classifiers. PART stands for partial decision tree algorithm which is a blend of C4.5 algorithm and RIPPER rule learning and works on the principle of separate-and-conquer. The algorithm produces sets of rules called 'decision lists' which are ordered set of rules. A new data is compared to each rule in the list in turn, and the item is assigned the category of the first matching rule (a default is applied if no rule successfully matches). PART builds a partial C4.5 decision tree in each iteration and makes the "best" leaf into a rule. Multiple back propagation algorithm (MBP) is a generalization of back propagation (BP) algorithm and has the network divided into two sub parts i.e space network and main network. The space network determines factors for the actuation of a neuron in the main network; therefore the neurons (in main network) get activated only for a set of training patterns instead for all as in multilayer perceptron. The learning of the network is achieved by adjusting the weights of the network (space + main) by Gradient descent approach as in back propagation. The Radial basis function (RBF) employs a non-liner function in the hidden layer and linear approach at the output layer. The Training and learning of RBF is very fast and also produces a good interpolation results. Decision table transforms the training dataset into a decision table and the test data is categorized into one of the possible classes by tracing the tuple in the decision table that fits the non-class attributes. For finding a proper attribute subset for inclusion in the table, it uses the wrapper method.

Each of the algorithms used for building the Ensemble were trained with the same data derived out of the database build in this work. Out of the database, multiple training and testing datasets were prepared which were used for validating the algorithms against 10 folds cross validation. Once all the individual algorithms were trained, using the training datasets, test instances were fed for classification, their corresponding results were noted and errors were calculated. The final classification done by this hybrid ensemble is thus a cumulative classification as resulted by pooled competence of multiple standalone algorithms and ensemble techniques. Figure 4 above presents a graphical representation classification schema.

Implementation of algorithms and results

For implementation of the various algorithms WEKA workbench (version 3.6.10) was employed. WEKA is a machine learning workbench, which is written in Java and consists of a collection of data analysis algorithms useful for designing predictive models as solutions to the Real-world problems. The datasets were prepared for feeding it to the WEKA workbench, for which it was converted to the required Attribute-Relation File Format (ARFF) format. ARFF format is an ASCII text file format that describes a list of cases sharing a set of attributes. Out of the 8091 Pap smear image instances present in the database, 5,664 (~70 %) instances were kept for training and randomly selected 2,427 (~30%) instances were kept for testing the system. Once the classification results for the test data from all the algorithms were obtained, they were arranged in a grid. This was followed by obtaining a model value for each of the test instance. The modal value represents the value that appears most often among a set of data values. This was achieved by writing a MATLAB code that scanned the results of all the implemented algorithms for the test data, and for each instance, produced the most frequently obtained classification result. These modal values were also inserted into the grid and the various performance metrics were obtained.

The correct classification of Pap smear images is most vital for precise screening and accurate diagnostic conclusions of cervical cancer. The automated classification of Pap smear images becomes complex problem using individual algorithms working alone; as such their classification efficiency is not at par. In this work we have effectively overcome this problem, and designed an ensemble system which significantly increases the classification results and as such the screening potential. The Hybrid ensemble technique so developed produces an increased efficiency of about 86% which is nearly 8% and 17% better as compared with the best and worst individual performers respectively. Among the algorithms used to build the ensemble, Rotation Forest algorithm depicted an efficiency of about 78.0%

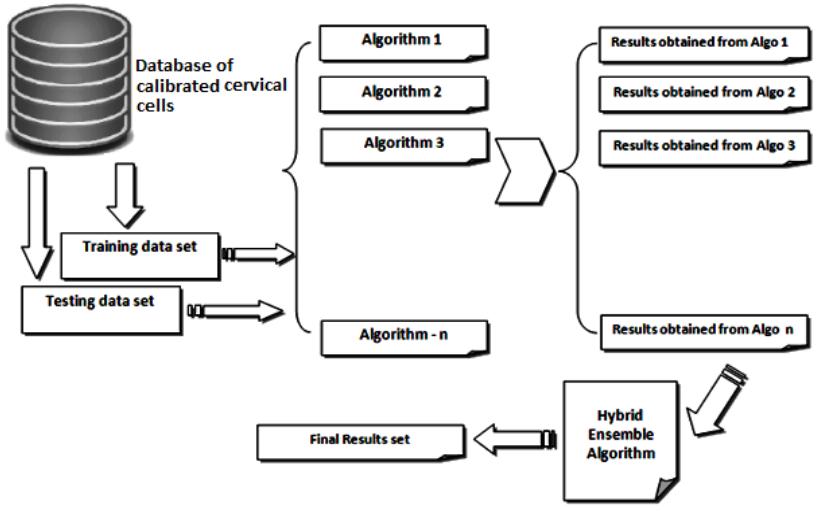

Figure 4: Block diagram representing the overall working of hybrid ensemble.

S No.	Artificial intelligence based Prediction model	2-Class Problem				Multiclass Problem			
		Correctly classified (%)	ROC Area	True Positive Rate	False Positive Rate	Correctly classified (%)	ROC Area	True Positive Rate	False Positive Rate
1	Bagging	95.2536	0.989	0.953	0.051	76.4629	0.942	0.765	0.081
2	PART	94.4083	0.974	0.944	0.066	73.5587	0.877	0.736	0.087
3	Decision Table	94.5384	0.986	0.945	0.065	74.0355	0.925	0.74	0.088
4	Decorate	95.0585	0.987	0.951	0.056	76.4629	0.939	0.769	0.077
5	END	93.9965	0.943	0.940	0.072	75.8778	0.938	0.759	0.082
6	Filtered Classifier	94.7334	0.964	0.947	0.056	74.5340	0.886	0.745	0.086
7	J48 Graft	94.4083	0948	0.944	0.066	73.6238	0.867	0.736	0.088
8	Multi-Class Classifier	94.9502	0.990	0.950	0.057	76.2029	0.937	0.762	0.082
9	Multiple Backpropagation ANN	95.4270	0.988	0.954	0.052	75.1409	0.928	0.751	0.084
10	Naïve Bayes	93.6064	0.975	0.936	0.076	69.7226	0.918	0.697	0.085
11	Radial Basis Function	94.1482	0.954	0.941	0.066	74.7941	0.921	0.748	0.090
12	Random Committee	95.3836	0.986	0.954	0.054	76.2029	0.930	0.762	0.081
13	Random Forest	95.2536	0987	0.953	0.057	75.3576	0.930	0.754	0.085
14	Random Sub Space	95.2536	0.989	0.953	0.053	76.9181	0.943	0.766	0.082
15	Rotation Forest	95.6220	0.990	0.956	0.049	78.0668	0.944	0.781	0.075
16	Novel Hybrid Ensemble	98.5700	0.998	0.988	0.021	86.6829	0.996	0.866	0.034

Table 4: Summary of 10 folds cross validation results for different algorithms.

followed by Random Sub Space (about 76.9%), Bagging (about 76.4%), Decorate (about 76.4%), Multiclass classifier (about 76.2%), Random Committee(about 76.2%), END (about 75.8%), Random forest (about 75.3%), Multiple back propagation artificial neural network (about 75.1%), Radial Basis Function (about 74.7%), FC (about 74.5%), Decision Table (about 74.0%), J48 graft (about 73.6%), PART (about 73.5%) and lastly by Naïve Bayes (about 69.7%). The results obtained are summarized below in Table 4.

For two class problem, the Hybrid ensemble technique showed an efficiency of about 98.8% which is nearly 3% and 5% better than the best and worst individual performers. For multiclass problem, the sensitivity and specificity for the hybrid ensemble were found to be 0.866 and 0.966, while as the false positive rate and false negative rate were found to be 0.034 and 0.134 respectively. The overall accuracy of the system for two class problem and multiclass problem are 0.9857 and 0.8668 respectively. Authors also compared the prognostic ability of the novel hybrid ensemble system with the machine learning based diagnostic systems available for Prostate cancer. Olivier Regnier-Coudert et al. in 2012 have done a comparison of 3 different machine learning models for prognosis of Prostate cancer [33]. The models used by them were Artificial neural networks, Bayesian Networks and Logistic Regression which expressed the ROC Areas of 0.656, 0.679 and 0.675 respectively. Compared with the prognostic results of these automated systembuild for prostate cancer, the novel hybrid ensemble system shows much better ROC Area of 0.996, which is very close to the best possible classification potential i.e. 1, by any classifier. As demonstrated by the results, the proposed hybrid ensemble method does more accurate screening which is precise and closer to actual diagnosis done by human medical expert, as compared to the other algorithms applied independently. The main advantage of the new hybrid ensemble system is proficient fusion of the screening potentials of various artificial intelligence based classification algorithms, thus increasing the overall predictive performance in segregating the abnormal Pap smear images from the normal ones.

Conclusion

Cervical cancer is the fourth most common cancer in women and the seventh overall, in the world. It is the one of the leading causes of female mortality due to cancer in the world. As reported by WHO the toll of cervical cancer in 2012 was 528000 with the death toll of 266000. A large majority of these cases (around 84%) are reported from the developing countries and under developed countries, as compared to the developed countries. This is attributed to mass level screening programs along with follow-ups at regular intervals done by the developed counters. Unfortunately in low income countries and countries having large population such mass level screening programs are very difficult to implement. Unfortunately, India is having more than 25% of whole world`s burden of cervical cancer, accounting for death of 8 patients every hour, about 75,000 patients every year. There is no health policy in India ,unlike US for mass cancer screening ,where regular Pap smear based cervical cancer screening is responsible for reduction in mortality by about 75% along with proper follow ups at regular intervals. Techniques that enable efficient screening of cervical cancer can potentially help in decreasing the incidence of cervical cancer and the concerned mortality rate. Database and other hybrid computer based technique that facilitate efficient automated screening of cervical cancer can aid in timely identification of cases having premalignant transformations that may later on develop into cervical cancers. Incorporating intelligent computer programs in cervical cancer diagnosis can assist and contribute a lot in mass level screening especially in low income countries. This way the overall burden over the world and that on the individual nations can be effectively mitigated. The digital database of cervical cells developed in this work precisely represents all the features that play a vital role in prognosis of cervical malignancy cancer and also shows a commendable efficiency in training many artificial intelligence based machine learning algorithms for screening of cervical cancer. The hybrid ensemble technique proposed in this work also shows good classification efficiency and thus has a potential to assist medical professionals and trained health care workers. This will consequently increase the diagnostic sensitivity and specificity of detection of cervical cancer by Pap smear screening.

The Digital Database and Ensemble Technique proposed in this work can be combined together to device a hardware tool which could facilitate easy and convenient mass level screening of cervical cancer.

The tool can be modeled in such a way as it would only required to input a set parameters and after processing using output the prognostic results. Such tool could be devised in such a way as it could be usable not only to the medical doctors but also to the trained medical technicians, for preliminary screening which would be subjected to further diagnostic subroutines if desired. This way the database and the hybrid ensemble technique would potentially help in mitigating not only the mortality due to cervical cancer but also would help in reducing the financial burden on individual nations and the world as a whole.

References

1. Arbyn M, Anttila A, Jordan J, Ronco G, Schenck U, et al. (2010) European Guidelines for Quality Assurance in Cervical Cancer Screening. Second edition--summary document. Ann Oncol 21: 448-458.
2. Lisboa PJ, Taktak AF (2006) The use of artificial neural networks in decision support in cancer: a systematic review. Neural Netw 19: 408-415.
3. Catto JW, Linkens DA, Abbod MF, Chen M, Burton JL, et al. (2003) Artificial Intelligence in Predicting Bladder Cancer Outcome: A Comparison of Neuro-Fuzzy Modeling and Artificial Neural Networks. Clin Cancer Res 9: 4172-4177.
4. Anagnostou T, Remzi M, Lykourinas M, Djavan B (2003) Artificial neural networks for decision-making in urologic oncology. Eur Urol 43: 596-603.
5. Bratko I, Kononenko I (1987) Learning Rules from Incomplete and Noisy Data. In: Phelps B (ed.) Interactions in Artificial Intelligence and Statistical Methods, Technical Press, Hampshire, England.
6. Lesmo L, Saitta L, Torasso P (1983) Fuzzy Production Rules: a Learning Methodology, Advances in Fuzzy Sets, Possibility Theory and Applications 181-198.
7. Catlett J (1991) On changing continuous attributes into ordered discrete attributes, Proceedings of European Working Session on Learning, Portugal, 164-178.
8. Clark P, Boswell R (1991) Rule Induction with CN2: Some Recent Improvements. Proceedings of European Working Session on Learning, Portugal, 151-163.
9. Nunez M (1990) Decision Tree Induction Using Domain Knowledge. Current Trends in Knowledge Acquisition, Amsterdam, IOS Press.
10. Hojker S, Kononenko I, Juka A, Fidler V, Porenta M (1988) Expert System's Development in Management of Thyroid Disease, proc. European Congress for nuclear medicine, Milano.
11. Horn KA, Compton P, Lazarus L, Quinlan JR (1985) an Expert System for Interpretation of Thyroid Assays in Clinical Laboratory, the Australian Computer Journal 17: 7-11.
12. Kern J, Dezelic G, Tezak-Bencic M, Durrigl T (1990) Medical Decision Making Using Inductive Learning Program, Proceedings of 1st Congress on Yougoslav Medical Informatics, Beogard.
13. Chaturvedi A, Gillison ML (2010) Human Papillomavirus and Head and Neck Cancer. Epidemiology, Pathogenesis, and Prevention of Head and Neck Cancer 87-116
14. Muñoz N, Bosch FX, de Sanjosé S, Herrero R, Castellsagué X, et al. (2003) Epidemiologic classification of human papillomavirus types associated with cervical cancer. N Engl J Med 348: 518-527.
15. World Health Organization (2006) Cancer.
16. GLOBOCAN (2008) summary table by cancer.
17. Shidham VB, Mehrotra R, Varsegi G, D'Amore KL, Hunt B, et al. (2011) INK4a immunocytochemistry on cell blocks as an adjunct to cervical cytology: Potential reflex testing on specially prepared cell blocks from residual liquid-based cytology specimens. CytoJournal 8: 1.
18. Kent A (2010) HPV Vaccination and Testing. Rev Obstet Gynecol 3: 33-34.
19. Saslow D, Solomon D, Lawson HW, Killackey M, Kulasingam SL, et al. (2012). American Cancer Society, American Society for Colposcopy and Cervical Pathology, and American Society for Clinical Pathology Screening Guidelines for the Prevention and Early Detection of Cervical Cancer. J Low Genit Tract Dis 16: 175-204.
20. Abid S, Vinod S (2013) Comparative analysis of machine learning techniques in prognosis of type II diabetes, AI & Society 29: 123-129.
21. Babak S, Siamak H, Ali DT (2014) A framework for diagnosing cervical cancer disease based on feedforward MLP neural network and ThinPrep histopathological cell image features. Neural Computing and Applications 24: 221-232.
22. Mat-Isa NA, Mashor MY, Othman NH (2008) An automated cervical pre-cancerous diagnostic system. Artif Intell Med 42: 1-11.
23. Chun-LC, Ming-YH (2009) The study that applies artificial intelligence and logistic regression for assistance in differential diagnostic of pancreatic cancer. Expert Systems with Applications 36: 10663-10672.
24. Genc-tav A, Selim A, Sevgen O (2012) Unsupervised segmentation and classification of cervical cell images. Pattern Recognition 45: 4151-4168.
25. Jantzen J, Norup J, Dounias G, Bjerregaard B (2005) Pap-smear benchmark data for pattern classification, Nature inspired Smart Information Systems (NiSIS) 1-9.
26. Nandakumar V, Kelbauskas L, Johnson R, Meldrum D (2011) Quantitative characterization of preneoplastic progression using single-cell computed tomography and three dimensional karyometry. Cytometry A 79: 25–34.
27. Bamford P, Lovell B (1996) A water immersion algorithm for cytological image segmentation, in Proc. APRS Image Segmentation Workshop, Sydney, Australia. 75–79.
28. Lezoray O, Cardot H (2002) Cooperation of color pixel classification schemes and color watershed: A study for microscopic images. IEEE Trans Image Process 11: 783–789.
29. Metzler V, Bienert H, Lehmann T, Mottaghy K, Spitzer K (1999) A novel method for quantifying shape deformation applied to biocompatibility testing. ASAIO J 45: 264-271.
30. Braumann UD, Kuska JP, Einenkel J, Horn LC, Lof- fler M, et al. (2005) Three-dimensional reconstruction and quantification of cervical carcinoma invasion fronts from histological serial sections. IEEE Trans Med Imaging 24: 1286–1307.
31. Abid S, Vinod S, Rajeev G (2015) Hybrid ensemble learning technique for screening of cervical cancer using Papanicolaou-smear image analysis. Personalized Medicine Universe 4: 54-62.
32. Das L, Sarkar T, Maiti AK, Naskar S, Das S, et al. (2014) Integrated cervical smear screening using liquid based cytology and bioimpedance analysis. J Cytol 31: 183-188.
33. Regnier-Coudert O, McCall J, Lothian R, Lam T, McClinton S, et al. (2012) Machine learning for improved pathological staging of prostate cancer: A performance comparison on a range of classifiers. Artif Intell Med 55: 25–35.

Citation: Sarwar A (2016) Artificial Intelligence Based Semi-automated Screening of Cervical Cancer Using a Primary Training Database. Cervical Cancer 1: 105. doi: 10.4172/2475-3173.1000105