Manuscript Details

Manuscript number	TISSUEANDCELL_2019_443_R1
Title	A comprehensive study on the multi-class cervical cancer diagnostic prediction on pap smear images using a fusion-based decision from ensemble deep convolutional neural network
Article type	Full Length Article

Abstract

The diagnosis of cervical dysplasia, carcinoma in situ and confirmed carcinoma cases is more easily perceived by commercially available and current research-based decision support systems when the scenario of pathologists to patient ratio is small. The treatment modalities for such diagnosis rely exclusively on precise identification of dysplasia stages as followed by The Bethesda System. The classification based on The Bethesda System is a multiclass problem, which is highly relevant and vital. Reliance on image interpretation, when done manually, introduces inter-observer variability and makes the microscope observation tedious and time-consuming. Taking this into account, a computer-assisted screening system built on deep learning can significantly assist pathologists to screen with correct predictions at a faster rate. The current study explores six different deep convolutional neural networks- Alexnet, Vggnet (vgg-16 and vgg-19), Resnet (resnet-50 and resnet-101) and Googlenet architectures for multi-class (four-class) diagnosis of cervical pre-cancerous as well as cancer lesions and incorporates their relative assessment. The study highlights the addition of an ensemble classifier with three of the best deep learning models for yielding a high accuracy multi-class classification. All six deep models including ensemble classifier were trained and validated on a hospital-based pap smear dataset collected through both conventional and liquid-based cytology methods along with the benchmark Herlev dataset.

Keywords	cervical dysplasia; deep learning; convolutional neural network; pap smear

Corresponding Author	Lipi B. Mahanta
Corresponding Author's Institution	Institute of Advanced Study in Science and Technology
Order of Authors	elima hussain, Lipi B. Mahanta, Chandana RayDas, Ratna Kanta Talukdar

Submission Files Included in this PDF

File Name [File Type]

covering letter of REVISED sibmission.docx [Cover Letter]

Checklist.doc [Checklist]

Response to reviewers.docx [Response to Reviewers]

Highlights.docx [Highlights]

revised paper - 31-01-2020-final.docx [Manuscript File]

Author statement.docx [Author Statement]

Data in brief - FINAL.docx [Data in Brief]

To view all the submission files, including those not included in the PDF, click on the manuscript title on your EVISE Homepage, then click 'Download zip file'.

Research Data Related to this Submission

There are no linked research data sets for this submission. The following reason is given:
The data has been co-submitted to Data in Brief via this submission

To,
The Editor,
Tissue and Cell Dated : 31-01-2020

Sub: Submission of REVISED Manuscript for publication

Dear Sir,

Please find the revised submission of the paper entitled "**A comprehensive study on the multi-class cervical cancer diagnostic prediction on pap smear images using a fusion-based decision from ensemble deep convolutional neural network**".

The title has been slightly changed to make it more meaningful. We are grateful for the valuable suggestions of the reviewers for improving the study and its documentation. We have tried to address all the points they have mentioned and believe that it is comprehensively done now.

We will be grateful if it is now considered for publication after the due process of the journal.

Thanking You,
Dr. Lipi B. Mahanta (corresponding author)
Associate Professor - II
Central Computational and Numerical Sciences Division
Institute of Advanced Study in Science & Technology (IASST)
(An Autonomous R&D Institute under Department of Science &Technology)
Vigyan Path, Paschim Boragaon, P.O.- Garchuk, Guwahati - 781035 (Assam), INDIA
Official web page : http://iasst.gov.in/dr-lipi-b-mahanta/
Email: lbmahanta@iasst.gov.in; lipimahanta@yahoo.co.in

Tissue and Cell – Validation report for animal experiments - *MANDATORY*

Authors reporting experiments on animals to *Tissue and Cell* and are requested to follow the ARRIVE guidelines during manuscript preparation. The following checklist adapted from Kilkenny C. et al, PLoS Biol. 2010 Jun 29;8(6):e1000412 is mandatory for authors to complete and upload at submission and will be used for initial screening of manuscripts. Not all fields might be applicable for all type of studies.

		Specify whether information is included in the manuscript (YES/NO)
Did your study involve animals and/or experiments on animals? MANDATORY		***YES/NO*** *(if NO, no need to complete the form further)*
Ethical statement MANDATORY	Indicate the nature of the ethical review permissions, relevant licenses (e.g. Animal [Scientific Procedure} Act 1986), and national or institutional guidelines for the care and use of animals, that cover the research	*No*
Study Design	The number of experimental and control groups	
	The experimental unit (e.g. a single animal, group, etc)	
	Details of how animals were allocated to experimental groups, including randomisation or matching if done	
	Any steps taken to minimise the effects of subjective bias when allocating animals to treatment (e.g. randomisation procedure) and when assessing results (e.g. if done, describe who was blinded and when)	
Sample Size	Total number of animals used in each experiment, and the number of animals in each experimental group	
	How number of animals was arrived at. Details of any sample size calculation used	
	Number of independent replications of each experiment, if relevant	
Experimental animals	Details of the animals used, including species, international strain nomenclature, sex, developmental stage (e.g. mean or median age plus age range) and weight (e.g. mean or median weight plus weight range)	
	Further relevant information such as the source of animals, genetic modification status (e.g. knock-out or transgenic), genotype, health/immune status, drug or test naïve, previous procedures, etc.	
Experimental procedures Info for each experiment and each experimental group (including control)	How (e.g. drug/substance formulation and dose, site and route of administration, anaesthesia and analgesia used [including monitoring], surgical procedure, method of euthanasia. Details of any specialist equipment used, including suppliers)	
	When (e.g. time of the day)	
	Where (e.g. home cage, laboratory, water maze)	
	Why (e.g. rationale for choice of specific anaesthetic, route of administration, drug/substance dose used)	
	Order in which the animals in the different experimental groups were treated and assessed	
Housing and husbandry	Housing (type of facility e.g. specific pathogen free (SPF); type of cage or housing, bedding material, number of cage companions)	
	Husbandry conditions (e.g. breeding programme, light/dark cycle, temperature, type of food, access to food and water, environmental enrichment)	
	Welfare-related assessments and interventions that were carried out prior to, during, or after the experiments	
Experimental outcomes	Clear definition of primary and secondary experimental outcomes assessed (e.g. cell death, molecular markers, behavioural changes)	
Statistical methods	Details of the statistical methods used for each analysis	
	Unit of analysis for each dataset (e.g. single animal, group of animals, single cell)	
	Method used to assess whether the data met the assumptions of the statistical approach	

Response to Decision Letter- Revise: 14 January, 2020

Ref: TISSUE AND CELL_2019_443

The authors would like to express their gratitude for pointing out the shortcomings in the manuscript and would like to re-submit with the following response and corrections. We have changed the title slightly to make it more meaningful. The title is now "**A comprehensive study on the multi-class cervical cancer diagnostic prediction on pap smear images using a fusion-based decision from ensemble deep convolutional neural network**".

The detailed response is as follows:

Reviewer 1

- **Comment 1:** Introduction: The first time you use an acronym, add its meaning. TBS, LBC, SCC. E.g.: LBC (Liquid-Based Cytology)

Response 1: We have done the following modifications by adding the meaning prior acronym in use as pointed out:

1. In Section 1. Introduction, page no. 1, paragraph 1, line no. 3 and line no. 4, we have added the meaning of previously used acronym as The Bethesda System (TBS) and liquid-based cytology (LBC).
2. In Section 1. Introduction, page no. 2, paragraph 3, line no. 1, we have added the meaning of previously used acronym as a low-grade squamous intra-epithelial lesion (LSIL) and, (2) high-grade squamous intra-epithelial lesion (HSIL).
3. In Section 1. Introduction, paragraph 3, page no. 2, line no. 4, we have added the meaning of previously used acronym as squamous cell carcinoma (SCC).
4. In Section 1. Introduction, paragraph 1, page no. 3, line no. 3, we have added the meaning of previously used acronym as the convolutional neural network (CNN).

- **Comment 2:** Add legends to the tables, explaining the meaning of the table and of the acronym: each figure, each table, must be able to be read by itself without using the text.

Response 2:

1. In Section 3. Materials and Methodology, sub-section 3.1 Image acquisition intended for generating the image datasets, page no. 5, table no. 1, the following legends have been included as Details of data acquired for the study; NILM: Negative for Intra-epithelial malignancy; LSIL: low grade squamous intra-epithelial lesion; HSIL: high grade squamous epithelial lesion; SCC: squamous cell carcinoma.

- **Comment 3:** Expand Discussion section: you must recall what you have obtained: using an ensemble classifier accuracy of 98.81% was reached, where three deep CNN, Resnet(s) & Googlenet reached the best performance.

Response 3:

1. We have included the section 5. Discussion on page no. 11 in this revised manuscript to highlight the proposed method and pointed out the main observations.
2. We have modified the section 6. Conclusion on page no. 11 since the revised manuscript included few experimental observations which were not included earlier and as such we have modified this section based on our findings.

Reviewer 2

- **Comment 1:** The manuscript fails to give a complete survey of the statistical behaviour and classification ability of both single (base) classifiers and the ensemble classifier. As an example, the rate of occurrence of false negative results is not analyzed in depth. However, this kind of misclassification has great relevance for biomedical applications, and often it is more prominent than false positive occurrence. In such a way, the paper seems a bit poor and should be improved before its publication.

Response 1: We have included Section 4. Results on page no. 7 to 10 with two new sub-sections: sub-section 4.1 Evaluation of the pre-trained classifier models on page no. 7 and sub-section 4.2 Evaluation of the multi-class classification task on page no. 8. The following new modifications are mentioned below:

1. In order to evaluate the classification ability of both single (base) classifiers and the ensemble classifier as pointed out, we tried to analyse the AUC **(**Area under the Curve**)** ROC **(**Receiver Operating Characteristics**)** curves of classifier models for every target classes. The AUC-ROC curve in fig. 2 and page no. 10 clearly highlights the superiority of ensemble classifier in terms of correct classification of individual classes with respect to false-positive and true positive rate as compared with alexnet, vggnet, resnet and googlenet. The description is given in in page no. 7, sub-section 4.2 Evaluation of the multi-class classification task.
2. The selection of three best suitable classifiers- resnet-50, resnet-101 and googlenet for combining ensemble classifier is done based on an assessment of individual fine-tuned classifier's (alexnet, vggnet, resnet and googlenet) performance during training, validation and testing phases. The performance evaluation is illustrated in sub-section 4.1 Evaluation of the pre-trained classifier models on page no. 7. The best result is highlighted in table no. 3, page no. 8 by googlenet followed by resnet-101 and resnet-50 in terms of the highest accuracy and lowest log-loss.
3. The rate of misclassification is considered by including false-positive rate and false-negative rate as performance metrics. Lower the value interpreted as a low misclassification rate. Fig. 1, page no. 9 in sub-section 4.2, highlights the superiority of the proposed ensemble method compared to other classifier models in terms of the highest accuracy, precision, recall, specificity and lowest false positive and false negative rate.

- **Comment 2:** I think you should adopt a more reliable definition of classifier accuracy, as well as to expand the statistical analysis of classifiers' performance (including additional statistical parameters, not necessarily supplementary data/images).

Response 2:

1. In the page no. 6, sub-section 3.3 Performance metrics for evaluation of the classification task, we have included precision, recall, specificity, false positive, false negative rate and area under the curve as an additional performance metric for measuring the classification ability of different classifier models used in the experiment. Table 2 on page no. 6 and sub-section 3.3 shows the definition of different performance metrics. The statistical analysis based on the performance metrics is highlighted in fig 1 on page no. 9 of sub-section 4.2.
2. In the page no. 10, below fig. 2, we have included McNemar's test to analyse the statistical significance of the classifier models.

❖ **Comment 3:** Define initializations and acronyms utilized along with the paper at their first occurrence (e.g. LBC, HPV, ANN, SCC, and so forth). The adoption of this rule strongly improves the manuscript readability.

Response 3: We have done the following modifications by adding the meaning prior acronym in use as pointed out:

1. In Section 1. Introduction, paragraph 1, page no. 1, line no. 3 and line no. 4, we have added the meaning of previously used acronym as The Bethesda System (TBS) and liquid-based cytology (LBC).
2. In Section 1. Introduction, paragraph 3, page no. 2, line no. 1, we have added the meaning of previously used acronym as a low-grade squamous intra-epithelial lesion (LSIL) and, (2) high-grade squamous intra-epithelial lesion (HSIL).
3. In Section 1. Introduction, paragraph 3, page no. 2, line no. 4, we have added the meaning of previously used acronym as squamous cell carcinoma (SCC).
4. In Section 1. Introduction, paragraph 1, page no. 3, line no. 3, we have added the meaning of previously used acronym as a convolutional neural network (CNN).

❖ **Comment 4:** Re-write the last two paragraphs of the Introduction (pages 4, 5).

Response 4: The following modifications are done:

1. Section 1. Introduction, page no. 2, paragraph 2, line no. 1 and line no. 1 and line no. 9.
2. Section 1. Introduction, page no. 2, paragraph 4.
3. Section 1. Introduction, page no. 3, paragraph 1, line no. 2.
4. Section 1. Introduction, page no. 3, paragraph 2

❖ **Comment 5:** Page 9; the paragraph above Fig. 1: did you truly mean "true negative" at the end of the third line?

Response 5: This is misinterpreted by false-negative as true-negative. False-negative rate is considered in performance evaluation as already mentioned. Section 5 (Discussion) on page no. 11 is modified now.

Highlights:

1. Six different deep convolutional neural networks (Alexnet, Vggnet-vgg16 and vgg19, Resnet-resnet50 and resnet101 and Googlenet) are compared to find the best classification accuracy model for cervical multi-class prediction using pap-smear images.
2. An ensemble classifier is introduced by combining three best architectures for the same classification problem.
3. To the best of our knowledge, this is the first work doing a comparative assessment of deep learning models using pap-smear images for the four-class (multi-class) prediction following The Bethesda System- NILM (normal class), LSIL (pre-cancerous class), HSIL (pre-cancerous class) and SCC (cancer class).
4. Ensemble classifier with combined Resnet50, Resnet101 and Google net provided the best classification accuracy.

A comprehensive study on the multi-class cervical cancer diagnostic prediction on pap smear images using a fusion-based decision from ensemble deep convolutional neural network

Elima Hussain[1], Lipi B. Mahanta[1]*, Chandana Ray Das[2], Ratna Kanta Talukdar[2]

1. Central Computational and Numerical Sciences Division, Institute of Advanced Study in Science and Technology, Guwahati- 781035, Assam, India

2. Guwahati Medical College & Hospital, Guwahati, Assam, India

**Corresponding Author: Lipi B. Mahanta, email: lbmahanta@iasst.gov.in*

Abstract: The diagnosis of cervical dysplasia, carcinoma in situ and confirmed carcinoma cases is more easily perceived by commercially available and current research-based decision support systems when the scenario of pathologists to patient ratio is small. The treatment modalities for such diagnosis rely exclusively on precise identification of dysplasia stages as followed by The Bethesda System. The classification based on The Bethesda System is a multiclass problem, which is highly relevant and vital. Reliance on image interpretation, when done manually, introduces inter-observer variability and makes the microscope observation tedious and time-consuming. Taking this into account, a computer-assisted screening system built on deep learning can significantly assist pathologists to screen with correct predictions at a faster rate. The current study explores six different deep convolutional neural networks- Alexnet, Vggnet (vgg-16 and vgg-19), Resnet (resnet-50 and resnet-101) and Googlenet architectures for multi-class (four-class) diagnosis of cervical pre-cancerous as well as cancer lesions and incorporates their relative assessment. The study highlights the addition of an ensemble classifier with three of the best deep learning models for yielding a high accuracy multi-class classification. All six deep models including ensemble classifier were trained and validated on a hospital-based pap smear dataset collected through both conventional and liquid-based cytology methods along with the benchmark Herlev dataset.

Keywords: cervical dysplasia; classification; deep learning; convolutional neural network; pap smear

1. Introduction

Pap smear test is considered as a renowned periodic screening tool for the detection of cervical pre-cancerous lesions or premalignant cells based on detailed microscopic observation. Diagnosis of cervical cancer is reliant on the Pap test by means of either conventional method or liquid-based cytology which is recommended by a clinician on subjective clinical assessment. An ideal Pap test report identifies the degree of malignancy if any and then confirms the classification category based on The Bethesda System (TBS) (Nayar and Wilbur, 2017) related to cervix cancer. It has been established that the liquid-based cytology(LBC) Pap test is more efficient and convenient than the conventional method (Cheung et al., 2003; Massad et al., 2001; Zhu et al., 2007) due to the mere fact that LBC can produce a cleaner and uniform slide for microscopic observations than the conventional method. LBC technique facilitates the breaking down of all heavier molecular particles like blood and mucus in the specimen with the help of centrifugation which confers a uniform slide at the end. LBC based results give better identification of cervical transformation zone cell-level components provided that the collected vial samples can also be used for the Human Papilloma Virus (HPV) testing. However, the option of using any of the screening tools does not produce any change in diagnosis results but might

be important for automating the overall screening system to assist a pathologist with disease diagnosis. Conventional images have more debris like red blood cells, inflammatory cells, etc. which have to be dealt with subsequent image pre-processing steps that may not be required for LBC images.

With the advent of artificial intelligence in the health care domain, predictions made by a decision support system can tackle issues on observer biases. A Pap test is used as a cervical cancer screening tool to detect squamous intraepithelial lesions or malignant growth if any. The Pap test report can exemplify early detection of the squamous intraepithelial lesion or cervical dysplasia where a rapid treatment plan can prevent its further development into invasive cancer. While prognosis based on Pap test report states that 57% of confirmed low grade squamous intraepithelial lesion (LSIL) cases regress to normal but more than 32% also progress to high grade squamous epithelial lesion (HSIL) or carcinoma in situ and almost 12% to invasive carcinoma (Maniar and Wei, 2017). With a limited skewed ratio of patients to pathologists, this screening test takes time to analyze the slides where a majority of cases are often confirmed as normal and without any dysplastic changes or sometimes even one cell with LSIL or HSIL characteristic confirmed its belonging to the LSIL and HSIL class. That is why the screening requires rigorous observation of individual cell characteristics for a multi-class diagnosis which is why automated prediction holds its significant relevance. There is yet another issue in most of the developing and underdeveloped countries where mass screening and awareness campaigns are dependent on cost-effective health resources with additional skilled manpower. Commercialized FDA-approved cervical screening systems do exist like the Focal Point GS Imaging System by BD (Becton Dickinson) (Wilbur et al., 2009)and Thin Prep Imaging system by HOLOGIC, Inc. (Biscotti et al., 2005) particularly for cervical cancer diagnosis but in countries like India, such systems may not be feasible because their high cost and maintenance are not effective during mass screening or even rural-based health check-ups. Consequently, this comprehensive study based on pap smear images holds strong in putting forward few recent deep learning networks for automated classification of cervical dysplasia into multiple classes which may assist pathologists in disease quantification and early detection where artificial intelligence may play a significant role and finally in planning abrupt prognosis treatment.

According to The Bethesda System (Nayar and Wilbur, 2017) cervical cell classification, there are two types of squamous epithelial lesions or premalignant lesions prior to normal class or negative for intraepithelial malignancy (NILM): (1) low-grade squamous intraepithelial lesion (LSIL) and, (2) high-grade squamous intraepithelial lesion (HSIL). In the case of LSIL class, dysplastic changes in the nuclear morphometry are observed as just starting off phase because of which they come under mild dysplasia category. Unlike LSIL, HSIL class cells have an abnormal nuclear size which is three or more times enlarged than normal class cells. Apart from nuclear enlargement, there are numerous cytological descriptions related to LSIL, HSIL, and squamous cell carcinoma (SCC) classes which are discussed in the reference book (Gray and Kocjan, 2010).

While automated diagnosis using machine learning algorithms rely on handcrafted features, deep learning methods can provide end-to-end classification by visual learning vast complex dimensional features without selective handcrafted feature engineering. The current study provides whole slide image analysis without relying on previous pipelined methods such as segmentation algorithms followed by feature extraction, and feature selection methods. This study using deep learning resulted in an automated diagnostic prediction of cervical cancer class from normal or healthy, dysplasia and carcinoma samples which may contribute to reducing observer biases or even sometimes manual workload of a pathologist.

Contributions of this paper can be summarized as:

(1) Development of an ensemble classifier that outperforms the classification accuracy yielding best result from selective deep learning models using majority voting technique (three best models selected under our problem domain).
(2) The current study is emphasized mainly on cervical dysplasia sub-types which is to our best of knowledge the first work doing ensemble classification on multi-class diagnosis using deep convolutional neural networks (CNN) with an inclusive study of six different deep CNN models using pap smear image analysis.
(3) We aim to reduce misclassification error by generalizing the ensemble classifier with real-world clinical pap smear datasets as well as validating the performance on a benchmark dataset.

This paper is organized as follows. Section **2** describes the prior art related to automated diagnosis on cervical dysplasia using machine learning or deep learning. Section **3** presents the experimental methodology and Section **4** and **5** highlights the results and discussion of our tentative findings. We conclude in Section **6.**

2. Prior art

Literature related to an automated cervical cancer diagnosis can be categorized into (1) study dataset whether cell level or smear level (whole slide image) and (2) whether classification persuaded using machine learning algorithms or deep learning algorithms. It is significant that smear level or whole slide image analysis will contribute to more rapid diagnosis reducing manual screening time which is why cell level dataset is not applied in the present study. Machine learning algorithms have been employed for the classification of cervical cancer but classification accuracy using such algorithms relies on a precise segmentation algorithm. Therefore, such segmentation algorithms configure into pipeline formats usually preceded with added numerous pre-processing steps which may destabilize classification output or may likewise increase misclassification rate. Few of the segmentation methods applied on cervical pap smear datasets and mentioned in the literature are radiating gradient vector flow (GVF) snake (Li et al., 2012), maximally stable extremal region (Bora et al., 2017), multi-scale hierarchical segmentation (Gençtav et al., 2012), morphological reconstruction and clustering(Plissiti et al., 2011)and global and local graph-cuts algorithm (Zhang et al., 2014). A few deep learning-based segmentation methods also exist in the same literature related to pap smear images (Song et al., 2015, 2014; Zhang et al., 2017b) but no further classification technique has been put forward using them. Litjens et. Al (Litjens et al., 2017) in his article have mentioned deep learning concepts on medical image classification and other tasks which is indeed becoming an emerging research field of medical imaging.

Work on multi-class diagnosis for pap smear images has been mentioned by Bora et. al (Bora et al., 2017), Marinakis et al. (Marinakis et al., 2008) and Changkong et al. (Chankong et al., 2014). Most of these conventional methods are able to efficiently remove clustered nuclei as well as unwanted debris like red blood cells and inflammatory cells leading to more accurate predictions. Bora et al. have mentioned about modified Maximally Stable Extremal Region (MSER) algorithm to specifically segment cellular artifacts. Changkong et al. have mentioned performance testing of five different classifiers – Bayesian classifier, linear discriminant analysis, k-nearest neighbor and artificial neural network out of which artificial neural network yielded best results for 2-class (binary) as well as 7-class (multi-class) prediction. Ensemble classifier using majority voting from three different classifier's decisions namely least square support vector machine, multi-layer perceptron, and random forest have been proposed by Bora et al. for NILM, LSIL, HSIL, and SCC diagnosis. Such methods are highly reliant on accurate segmentation output which only succeeds the rest of the pipeline thus making it slow and increasing the chance of classification error. With deep learning, the aforesaid multi-class

prediction or diagnosis may be incorporated without prior segmentation using only the smear level images. Most of these deep learning works on pap smear images deal with either binary classification problems or using only a single cell-level (cropped-out) dataset. Srishti et al. (Gautam et al., 2018) have reported an F-score of 0.90 on the benchmark dataset using a proposed patch-based CNN classifier to address the 2-class classification problem but applied only for the single cell-level pap smear dataset. Binary classification using such single-cell images has also been proposed by Zhang et al. (Zhang et al., 2017a) using Convnet architectures that have achieved 98.3% test accuracy on comparison with benchmark dataset. A deep convolutional neural network architecture namely a Deep-cerv network has been proposed for binary classification of pap smear images by Nirmaljith et al. (Jith et al., 2018) that has achieved 99.6% test accuracy as reported. Thus, from literature, it has been clear that no such deep learning models have been introduced so far for automated multi-class diagnosis of cervical cancer using the pap-smear images.

3. Materials and Methodology

3.1. Image acquisition intended for generating the image datasets

A hospital-based dataset of pap smear samples was collected to deal with real-world clinical setup. Developed algorithms hold their ground by comparing results with the publicly available benchmark datasets (Herlev University dataset for pap smear images). This will work as a base for the development of further new algorithms with module wise improvement that deals with real-world scenarios while keeping global standards within sight. A total of 1670 image datasets using the liquid-based cytology Pap test belonging to patients who came for cervical screening tests were collected from three distinguished medical diagnostic centers of the North-eastern regions of India namely Babina Diagnostic Pvt. Ltd, Imphal, Gauhati Medical College and Hospital, Guwahati and Dr B. Borooah Cancer Institute, Guwahati. Another 1320 images were collected using the conventional Pap test from AyursundraPvt. Ltd, Guwahati and Dr B. Borooah Cancer Institute, Guwahati. All slides were prepared following standard Pap staining protocol. Ethical permission was taken for the study from the institutional ethics board (Registration number ECR/248/Indt/AS/2015 of Rule 122DD, Drugs and Cosmetics Rule, 1945 of India) of the Institute of Advanced Study in Science and Technology, Guwahati [No. IEC. IEC(HS)/IASST/1082/2015-16/3]. All samples used in the study also involve appropriate patient consent from the respective diagnostic centers. The true class according to TBS cervical cell classification (NILM, LSIL, HSIL, and SCC) to which the pap smear images belong were first manually confirmed by an expert pathologist. This is used as ground truth labels in the study for confirming our experimental findings respectively. The images were captured using a Leica ICC50 high-definition microscope with 400× magnification. **Table. 1** gives an overview of the captured images in detail. The benchmark Herlev University dataset consists of seven classes, where superficial squamous, intermediate squamous and columnar squamous were grouped into the normal class or NILM, mild dysplasia grouped as LSIL class, moderate dysplasia and severe dysplasia as HSIL class and carcinoma in situ as SCC class. Herlev dataset is downloaded from the URL: http://labs.fme.aegean.gr/decision/downloads

3.2. Building a classifier model

Transfer learning comes into play when building a convolutional neural network from scratch is just impractical in medical research. This is because of a scarcely available clinical dataset and requisite computation resources. As an alternative, publicly available CNN models that are already pre-trained

with a natural image dataset can be fine-tuned with their own dataset in specific application areas to which transfer learning comes into account. Transfer learning refers to the fine-tuning of a pre-trained

Table. 1: Details of data acquired for the study

LBC dataset (own)	Images acquired
NILM	900
LSIL	360
HSIL	250
SCC	160
Total no. of images	**1670**
Conventional dataset (own)	
NILM	796
LSIL	247
HSIL including SCC	278
Total no. of images	**1320**
Herlev dataset(benchmark)	
Superficial squamous	74
Intermediate squamous	70
Columnar squamous	98
Mild dysplasia	182
Moderate dysplasia	146
Severe dysplasia	197
Carcinoma in situ	150
Total no. of images	**917**

NILM: Negative for Intraepithelial malignancy; LSIL: low grade squamous intraepithelial lesion; HSIL: high grade squamous epithelial lesion; SCC: squamous cell carcinoma.

network on a labeled large-scale natural image dataset. Fine-tuning involves adjusting the weights of the pre-trained network by continuing with backpropagation. Typically, apart from weight adjustment, fine-tuning also involves resetting or truncating the last fully connected layers, which can be viewed as a classification layer along with a smaller learning rate being applied to the pre-trained layers. The goal is to adapt deep features to the new datasets. More different is the latter from the original dataset, more parameters or layers must be reset. Traditional supervised machine learning paradigm breaks down due to limited labeled datasets to train a model whereas in solving similar problems, we can directly transfer the knowledge to our target domain.

Initially, publicly available pre-trained models namely, Alexnet (Krizhevsky et al., 2012), Vggnet (Vgg-16, Vgg-19) (Simonyan and Zisserman, 2015), Resnet (resnet-50, resnet-101) (He et al., 2016) and Googlenet (Szegedy et al., 2015) were used as candidate training models for pap smear images. The model selection was made based on their well-known performance for different classification tasks. Accordingly, transfer learning was used for classifying the cervical classes- NILM, LSIL, HSIL, and SCC. Since this is a four-class classification task, the last fully connected layers for each pre-trained models were replaced with a modified fully connected layer comprising of four output nodes representing these four classes. Model training was conducted using Adam optimizer for 30 epochs with a batch size of 32 images. For the pre-trained models including ensemble the hyper-parameters used were momentum being 0.9 with a weight decay of 0.0005 and a network learning rate of 0.001, which was decreased by a factor of 10 at every 10 epochs till the networks reach a convergence point. Network training was implemented using Keras package with python environment on a GPU based system having Intel®Core i7® 8750H processor with 6GB memory and GTX® 1060 graphics card.

3.2.1 Ensemble classifier

An Ensemble classifier scheme works on seeking a maximum number of classifiers' decisions and weighing their decisions at the same time to increase efficiency and performance of the final classification task. For this, the output of the six publicly available deep models namely, Alexnet, Vgg-16, Vgg-19, Resnet-50, Resnet-101, and Googlenet were evaluated based on performance and three best models (Resnet-50, Resnet-101, and Googlenet) were combined to generate an ensemble classifier using majority voting technique. The ensemble model chooses a class based on its highest number of votes received so far. Thus, the decision of i^{th} classifier can be defined as $D(a,b) \in (0,1)$ such that $a=1,2,3...,M$ and $b=1,2,3...,N,$ where M is the number of classifiers and N is the number of classes. In such cases if i^{th} classifier chooses class ω_b, then $D(a,b) = 1$ and 0 otherwise. An extensive analysis of majority voting can be found in (Polikar, 2006). The ensemble decision for majority voting can be described as follows:

$$\sum_{a=1}^{M} D(a,b) = \max_{b=1}^{N} \sum_{a=1}^{T} D(a,b) \quad (1)$$

3.3. Performance metrics for evaluation of the classification task

The prediction made by the classifiers can be evaluated using the following metrics meant for multi-class classification (Sokolova and Lapalme, 2009). Here, true-positives and true-negatives denote the number of positive and negative classes correctly predicted by individual classifiers; false-positive and false-negative denotes the number of positive and negative classes that are incorrectly predicted by individual classifiers. Accuracy increases the rate of true predicted classes.

Table. 2: Performance measures for multi-class classification. Here, for many classes, C_i,tp_i represents true-positive class; fp_i represents false-positive class; tn_irepresents true-negative class; fn_i represents false-negative class; l denotes total classes

Performance metrics	Formula	Focus
Average Accuracy	$\frac{\sum_{i=1}^{l} \frac{tp_i + tn_i}{tp_i + tn_i + fp_i + fn_i}}{l}$	Calculates average prediction efficiency of individual classifiers for each class
Precision or Positive Predictive Value	$\frac{\sum_{i=1}^{l} \frac{tp_i}{tp_i + fp_i}}{l}$	Agreement of class labels with positive labels given by individual classifiers
Recall or Sensitivity	$\frac{\sum_{i=1}^{l} \frac{tp_i}{tp_i + fn_i}}{l}$	Calculates individual classifier's efficiency for identifying positive labels. This gives the true positive rate.
Specificity	$\frac{\sum_{i=1}^{l} \frac{tn_i}{tn_i + fp_i}}{l}$	Calculates individual classifier's efficiency for identifying negative labels. This gives the true negative rate.
False-positive rate (or 1- Specificity)	$\frac{\sum_{i=1}^{l} \frac{fp_i}{tn_i + fp_i}}{l}$	Calculates individual classifier's efficiency for identifying false positive labels.
False-negative rate	$\frac{\sum_{i=1}^{l} \frac{fn_i}{tn_i + fn_i}}{l}$	Calculates individual classifier's efficiency for identifying false-negative labels.
Area Under Curve	$\frac{1}{2}\sum_{i=1}^{l}\left(\frac{tp_i}{tp_i + tn_i} + \frac{tn_i}{tn_i + fp_i}\right)$	Calculates individual classifier's effectiveness in avoiding false classification

4. Results

In this study, we evaluated and analyzed the performance of six different deep learning models using transfer learning and compared the ensemble classifier's performance with the individual pre-trained or fine-tuned models for cervical dysplasia multi-class diagnostic prediction. Classification using publicly available benchmark datasets (Herlev University dataset) and pap smear images collected through two different methods- liquid-based cytology and conventional pap test are evaluated. Results were verified with manually annotated images by an expert pathologist that served as ground-truth. This performance evaluation is presented by following steps mentioned below:

(a) Resize all raw pap smear images from three different datasets into a uniform dimension of 987×654pixels.
(b) Split the dataset into training, validation and test dataset based on the train-test split strategy (Mohanty et al., 2016). From a total of 3907 pap smear images, 3125 images were used for network training whereas 781 images were used for model validation from which again 316 images were used as test datasets.
(c) Train the classifiers (Alexnet, VggNet-16 and 19, Resnet-50 and 101, Googlenet and Ensemble) using a training dataset for prediction.
(d) Use the trained classifiers for predicting the test dataset.
(e) Analyze the performance metrics: average accuracy, precision, recall, specificity, false-positive and false-negative rate and area under the curve.

4.1 Evaluation of the pre-trained classifier models

An assessment of suitable classifiers from existing deep convolutional neural networks (Alexnet, vgg-16, vgg-19, resnet-50, resnet-101 and googlenet) pre-trained on ImageNet and fine-tuned for the task of cervical dysplasia prediction using pap-smear images was done. These networks require large datasets for optimal training and to overcome this, we have fine-tuned the pre-trained networks for our problem. Fine-tuning the individual models is carried out particularly as specified in **Section 3.2**. On the basis of the loss and accuracy consistency during training, validation and testing for the individual models, three classifier models showed the best performance. Results highlighted in Table. 3. prove that fine-tuned Googlenet followed by Resnet-50, and Resnet-101 converges better than Alexnet, Vgg-16, and Vgg-19.

Googlenet performed the classification task gaining the highest accuracy and lowest log-loss whereas Alexnet acquired the least accuracy and highest log-loss in the same classification problem. At epoch 30, only three classifier models – Resnet-50, Resnet-101 and Googlenet had accuracy above 90% consistently for liquid-based cytology, conventional and Herlev pap smear datasets with considerably minimized log-loss. Deeper networks like Googlenet perform better than shallower networks like Alexnet with a 37% reduced number of trainable parameters. This establishes that fine-tuned Googlenet, Resnet-50, and Resnet-101 can converge easily and can significantly learn more visual deep features of pap smear images than the other pre-trained models.

4.2 Evaluation of the multi-class classification task

The three best-performance classifiers were chosen for the ensemble classifier model. We validated the classification task with all six deep learning models as well as the ensemble model for a comparative assessment. For this, we performed receiver operating characteristic (ROC) analysis for fine-tuned Alexnet, Vgg-16, Vgg-19, Resnet-50, Resnet-101, Googlenet and the proposed ensemble classifier. Fig.

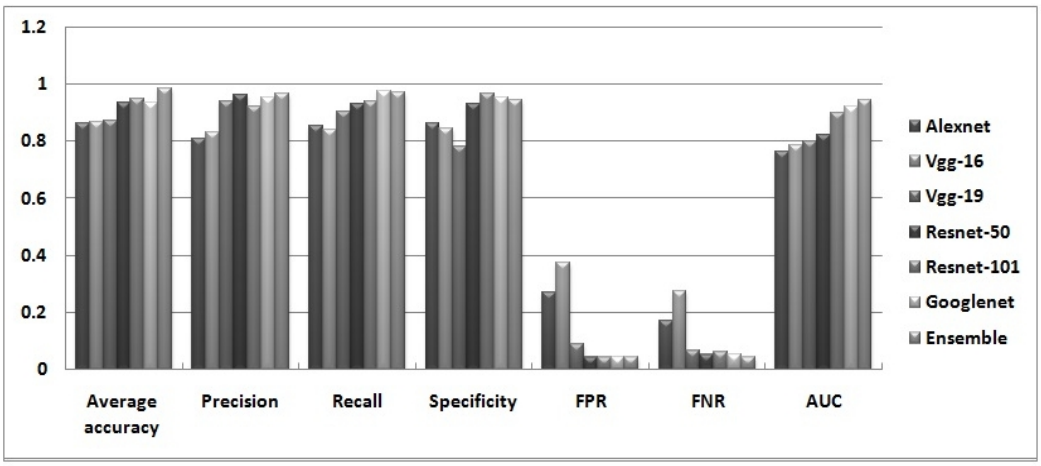
1 highlights the ROC curves for proposed and individual models considering all the multiple classes: normal or healthy (NILM), LSIL, HSIL and SCC cases. The area under the curve (AUC) was computed

Table. 3: Comparison of fine-tuned models using estimated accuracy and loss during training, validation, and testing. Bold values indicate the best performance classifier.

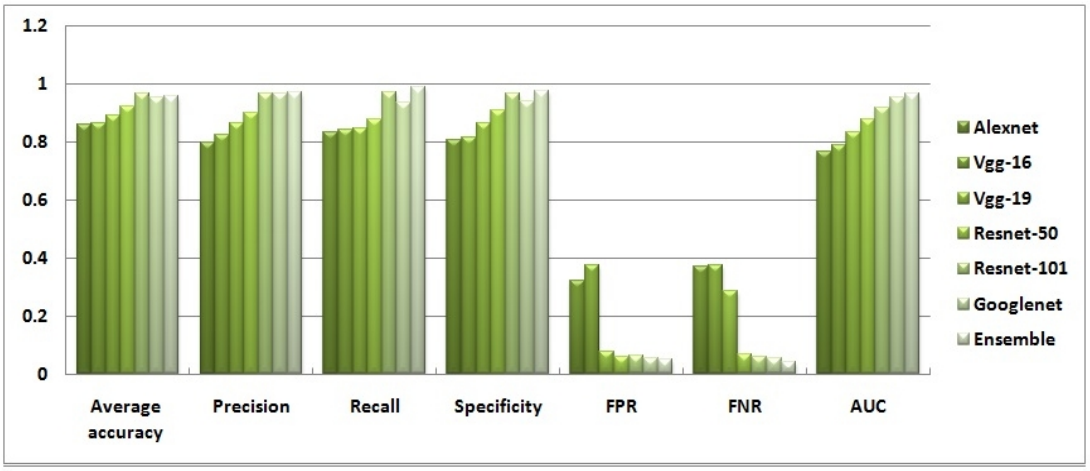
Datasets	**Fine-tuned models**	**No. of trainable Parameter (millions)**	**Training accuracy at epoch 30 (%)**	**Training loss at epoch 30**	**Validation accuracy at epoch 30 (%)**	**Validation loss at epoch 30**	**Testing accuracy at epoch 30 (%)**	**Testing loss at epoch 30**
Liquid based cytology	Alexnet	61	84.41	0.851	84.34	0.678	82	0.578
	Vgg-16	138	87.19	0.674	86.67	0.456	87.16	0.341
	Vgg-19	144	85	0.773	84.47	0.712	85.16	0.702
	Resnet-50	25.6	91	0.025	90.12	0.019	91.78	0.015
	Resnet-101	45.6	92.57	0.020	93	0.024	92.61	0.024
	Googlenet	23	95.45	0.011	96.67	0.014	95.12	0.014
Conventional	Alexnet	Same as above	80	0.567	79.67	0.671	82	0.670
	Vgg-16		87.43	0.709	88.89	0.666	87	0.545
	Vgg-19		86.14	0.145	86.45	0.146	87.33	0.122
	Resnet-50		91.13	0.023	93	0.023	92	0.021
	Resnet-101		94	0.028	92.67	0.013	94.89	0.015
	Googlenet		96.67	0.017	96	0.011	97.18	0.016
Herlev	Alexnet	Same as above	83	0.613	83.74	0.555	80	0.644
	Vgg-16		85.67	0.267	84.11	0.232	83.37	0.545
	Vgg-19		87.01	0.471	86.23	0.418	84.55	0.333
	Resnet-50		93.51	0.054	90	0.056	89.37	0.034
	Resnet-101		92.22	0.022	94.16	0.021	90.45	0.029
	Googlenet		96.17	0.015	97	0.015	95.67	0.015

considering a 97% confidence interval (Hanley et al., 1983). From Fig. 1 it is clearly seen that for each of the four classes, the proposed ensemble classifier achieved the highest AUC. Five-fold cross-validation was performed which involves dividing the data into five parts and fitting the model using 90% of the split data and then finally predicting with the remaining 10%. A comparison of different performance metrics of classification results is presented in Fig. 1 below.

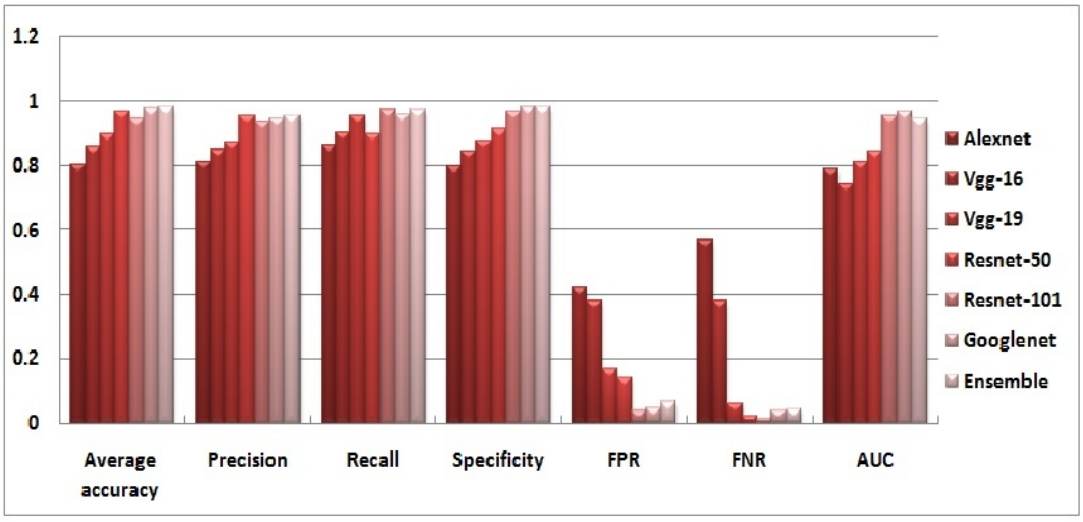
In order to visualize our clinical findings based on the multi-class classification task, we have checked the AUC **(**Area Under the Curve**)** ROC **(**Receiver Operating Characteristics**)** curve for the individual classifier models. AUC-ROC is a good performance metric used in diagnostic tests to analyze how well a classifier model is capable of correctly distinguishing normal or healthy patients (negative class) and actual patients suffering from a certain disease (positive class). In other words, the AUC-ROC illustrates the diagnostic decision by comparing the trade-off between true positive rate (TPR) and false-positive rate (FPR) as the two operating characteristics criterion. The false-positive case occurs when a healthy person is predicted wrongly to have a disease and its minimization is considered as an intuitive problem for any diagnostic test. For this, a good classification result is only reciprocated by the points lying above the diagonal line (random). Fig. 2 illustrates the ROC curve plotted with TPR against FPR for NILM, LSIL, HSIL and SCC classes. From the figure, it is clear that the ensemble classifier displays the best criterion achieving an AUC 0.97, which means 97% of the classifier model has been able to distinguish positive and negative class.



(a)



(b)



(c)

Fig. 1: Comparison of different performance metrics for evaluation of classifier models using (a)liquid-based cytology dataset, (b) conventional dataset and (c) Herlev dataset.

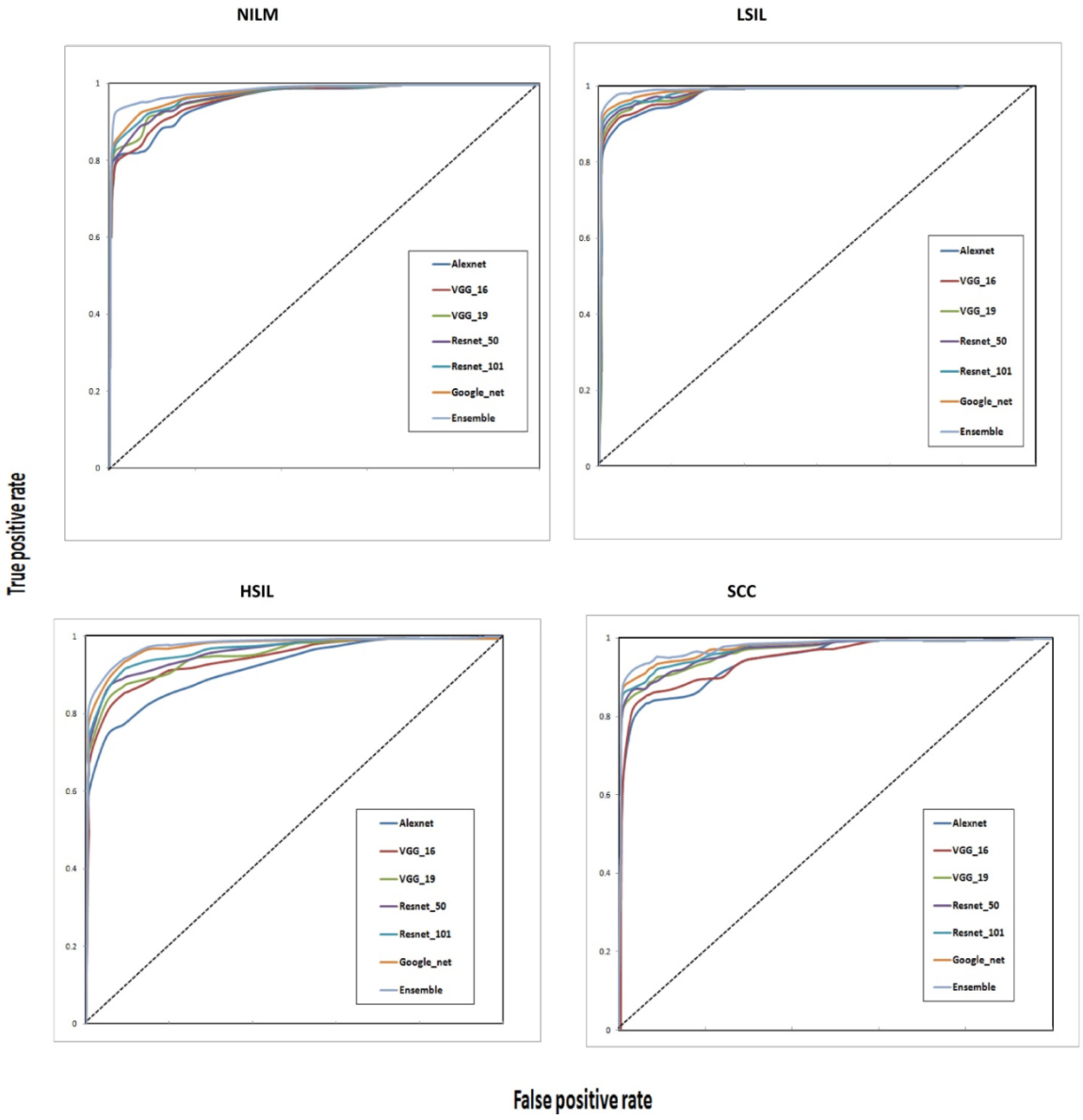


Fig. 2: Comparison of AUC-ROC curve across all experimental datasets

McNemar's test was conducted to confirm the statistical significance of the consistency of the classifiers against ground-truth by evaluating the number of classified and not classified instances of the individual class. This analysis confirmed the superiority of the proposed ensemble method to be statistically significant ($p< 0.01$) than the other networks. The findings are presented in the table below (Table. 4).

Table. 4: *p*-values of McNemar's test conducted for different networks and for the four classes.

Models	NILM	LSIL	HSIL	SCC
Alexnet	0.789	0.677	0.789	0.567
Vgg-16	0.555	0.321	0.677	0.578
Vgg-19	0.078	0.456	0.067	0.291
Resnet-50	0.058	0.056	0.057	0.067
Resnet-101	0.051	0.042	0.025	0.034
Ensemble	**0.01**	**0.01**	**0.00**	**0.01**

5. Discussion

From our findings, it can be considered that automatic multi-class prediction of cervical dysplasia can pave way for decision support systems that can help pathologists in disease quantification and prognosis treatment. From the experimental observations, it is also clear that the proposed ensemble classifier can predict the four target classes with high accuracy, precision and recall than other candidate classifier models. Based on our observations, we have summarized below the main observations to highlight the proposed method:

1. We have given emphasis to validate our diagnostic test by taking into account the rate of occurrence of false-positive and false-negative results. In our case, false negative is interpreted as patients with cervical dysplasia or carcinoma but the diagnostic test shows negative result whereas false-positive implies patients with cervical dysplasia but diagnosed with a positive result. The proposed ensemble classifier outperforms the other models with a low false-positive and false-negative rate.
2. The ensemble model can be considered as the most suitable and generalized model for cervical pap smear image classification since it is combined by three optimized base deep learning models in terms of training and testing accuracy and with the reduced number of trainable parameters. The proposed model is efficient in terms of AUC values which are more than 90% than Alexnet, Vggnet, Resnet and Googlenet.
3. This method can classify whole slide pap smear images without relying on segmentation techniques which makes the framework more robust.

6. Conclusion

Patients with low-grade and high-grade squamous intraepithelial lesions are at a high risk of progressing into cervical squamous cell carcinoma or even invasive carcinoma if not detected early. The screening test through pap smear allows routine examination of lesions by a pathologist. The present article explains a deep learning method for assisting a pathologist for automatic and rapid diagnostic prediction using pap smear image analysis. This method overcomes incorrect predictions and does not require segmentation and hand-engineered feature extraction steps, unlike other conventional methods. The proposed method is evaluated using three datasets: liquid-based cytology, conventional and Herlev datasets where the better result is reported by the ensemble classifier with 0.989 accuracy, 0.978 sensitivity and 0.979 specificity. Findings prove that this ensemble method is advantageous as the emphasis is given on all the stages of dysplasia and suggests potential utility for early-stage disease diagnosis. With an alarming growth of cervical cancer patients where the ratio of pathologists for disease diagnosis is very limited prior to screening patients, the suggested method may ease the overall screening protocol.

References

Biscotti, C. V., Dawson, A.E., Dziura, B., Galup, L., Darragh, T., Rahemtulla, A., Wills-Frank, L., 2005. Assisted primary screening using the automated ThinPrep Imaging System. Am. J. Clin. Pathol. 123, 281–287. https://doi.org/10.1309/AGB1MJ9H5N43MEGX

Bora, K., Chowdhury, M., Mahanta, L.B., Kundu, M.K., Das, A.K., 2017. Automated classification of Pap smear images to detect cervical dysplasia. Comput. Methods Programs Biomed. 138, 31–47. https://doi.org/10.1016/j.cmpb.2016.10.001

Chankong, T., Theera-Umpon, N., Auephanwiriyakul, S., 2014. Automatic cervical cell segmentation and classification in Pap smears. Comput. Methods Programs Biomed. 113, 539–556. https://doi.org/10.1016/j.cmpb.2013.12.012

Cheung, A.N.Y., Szeto, E.F., Leung, B.S.Y., Khoo, U.S., Ng, A.W.Y., 2003. Liquid-Based Cytology and Conventional

Cervical Smears: A Comparison Study in an Asian Screening Population. Cancer 99, 331–335. https://doi.org/10.1002/cncr.11786

Gautam, S., Bhaskar, A., Sao, anil k, K.K., H., 2018. CNN based segmentation of nuclei in PAP-smear images with selective pre-processing, in: Medical Imaging.

Gençtav, A., Aksoy, S., Önder, S., 2012. Unsupervised segmentation and classification of cervical cell images. Pattern Recognit. 45, 4151–4168. https://doi.org/10.1016/j.patcog.2012.05.006

Gray, W., Kocjan, G., 2010. Diagnostic cytopathology. Churchill Livingstone.

Hanley, J.A., Mcneil, J., Ph, D., 1983. A method of Comparing Operating Curves the Areas from Receiver Derived the Same Cases '. Radiology 148, 839–843.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. 2016-Decem, 770–778. https://doi.org/10.1109/CVPR.2016.90

Jith, O.U.N., Harinarayanan, K.K., Gautam, S., Bhavsar, A., Sao, A.K., 2018. DeepCerv: Deep Neural Network for Segmentation Free Robust Cervical Cell Classification, in: First International Workshop, COMPAY 2018, and 5th International Workshop, OMIA 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16 - 20, 2018, Proceedings.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. ImageNet Classification with Deep Convolutional Neural Networks, in: NIPS.

Li, K., Lu, Z., Liu, W., Yin, J., 2012. Cytoplasm and nucleus segmentation in cervical smear images using Radiating GVF Snake. Pattern Recognit. 45, 1255–1264. https://doi.org/10.1016/j.patcog.2011.09.018

Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., van der Laak, J.A.W.M., van Ginneken, B., Sánchez, C.I., 2017. A survey on deep learning in medical image analysis. Med. Image Anal. 42, 60–88. https://doi.org/10.1016/j.media.2017.07.005

Maniar, K.P., Wei, J.-J., 2017. Pathology of Cervical Carcinoma, in: The Global Library of Women's Medicine. International Federation of Gynecology and Obstetrics. https://doi.org/DOI 10.3843/GLOWM.10230

Marinakis, Y., Marinaki, M., Dounias, G., 2008. Particle swarm optimization for pap-smear diagnosis. Expert Syst. Appl. 35, 1645–1656. https://doi.org/10.1016/j.eswa.2007.08.089

Massad, L.S., Collins, Y.C., Meyer, P.M., 2001. Biopsy correlates of abnormal cervical cytology classified using the bethesda system. Gynecol. Oncol. 82, 516–522. https://doi.org/10.1006/gyno.2001.6323

Mohanty, S.P., Hughes, D.P., Salathé, M., 2016. Using Deep Learning for Image-Based Plant Disease Detection. Front. Plant Sci. 7, 1–10. https://doi.org/10.3389/fpls.2016.01419

Nayar, R., Wilbur, D.C., 2017. The bethesda system for reporting cervical cytology: A historical perspective. Acta Cytol. 61, 359–372. https://doi.org/10.1159/000477556

Plissiti, M.E., Nikou, C., Charchanti, A., 2011. Automated detection of cell nuclei in Pap smear images using morphological reconstruction and clustering. IEEE Trans. Inf. Technol. Biomed. 15, 233–241. https://doi.org/10.1109/TITB.2010.2087030

Polikar, R., 2006. Ensemble based systems in decision making. IEEE Circuits Syst. Mag. 6, 21–44. https://doi.org/10.1109/MCAS.2006.1688199

Simonyan, K., Zisserman, A., 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition, in: ICLR. pp. 1–14.

Sokolova, M., Lapalme, G., 2009. A systematic analysis of performance measures for classification tasks. Inf. Process. Manag. 45, 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

Song, Y., Zhang, L., Chen, S., Ni, D., Lei, B., Wang, T., 2015. Accurate segmentation of cervical cytoplasm and nuclei based on multiscale convolutional network and graph partitioning. IEEE Trans. Biomed. Eng. 62, 2421–2433. https://doi.org/10.1109/TBME.2015.2430895

Song, Y., Zhang, L., Chen, S., Ni, D., Li, B., Zhou, Y., Lei, B., Wang, T., 2014. A deep learning based framework for accurate segmentation of cervical cytoplasm and nuclei. 2014 36th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBC 2014 2903–2906. https://doi.org/10.1109/EMBC.2014.6944230

Szegedy, C., Vanhoucke, V., Shlens, J., 2015. Rethinking the Inception Architecture for Computer Vision, in: CVPR. pp. 2818–2826.

Wilbur, D.C., Black-Schaffer, W.S., Luff, R.D., Abraham, K.P., Kemper, C., Molina, J.T., Tench, W.D., 2009. The Becton Dickinson focal point GS imaging system: Clinical trials demonstrate significantly improved sensitivity for the detection of important cervical lesions. Am. J. Clin. Pathol. 132, 767–775. https://doi.org/10.1309/AJCP8VE7AWBZCVQT

Zhang, L., Kong, H., Chin, C.T., Liu, S., Chen, Z., Wang, T., Chen, S., 2014. Segmentation of cytoplasm and nuclei of abnormal cells in cervical cytology using global and local graph cuts. Comput. Med. Imaging Graph. 38, 369–380. https://doi.org/10.1016/j.compmedimag.2014.02.001

Zhang, L., Lu, L., Nogues, I., Summers, R.M., Liu, S., Yao, J., 2017a. DeepPap: Deep convolutional networks for cervical cell classification. IEEE J. Biomed. Heal. Informatics 21, 1633–1643. https://doi.org/10.1109/JBHI.2017.2705583

Zhang, L., Sonka, M., Lu, L., Summers, R.M., Yao, J., 2017b. COMBINING FULLY CONVOLUTIONAL NETWORKS AND GRAPH-BASED APPROACH FOR AUTOMATED SEGMENTATION OF CERVICAL CELL NUCLEI Radiology and Imaging Sciences Department, National Institutes of Health (NIH), Bethesda MD Iowa Institute for Biomedical Imaging and Dep. Isbi 406–409.

Zhu, J., Norman, I., Elfgren, K., Gaberi, V., Hagmar, B., Hjerpe, A., Andersson, S., 2007. A comparison of liquid-based cytology and Pap smear as a screening method for cervical cancer. Oncol. Rep. 18, 157–160. https://doi.org/10.3892/or.18.1.157

Elima Hussain: Data curation, Investigation, Methodology, Software, Writing- Original draft preparation

Lipi B. Mahanta: Conceptualization, Supervision, Data curation, Investigation, Formal analysis, Writing- Reviewing and Editing,

Chandana Ray Das: Resources, Validation.

Ratna Kanta Talukdar: Resources, Validation.

Article Title

Liquid based-cytology Pap smear dataset for automated multi-class diagnosis of pre-cancerous and cervical cancer lesions.

Authors

Elima Hussain[1], Lipi B. Mahanta[1*], Himakshi Borah[2], Chandana Ray Das[2]

Affiliations

[1.]Central Computational and Numerical Sciences Division, Institute of Advanced Study in Science and Technology, Guwahati, Assam, India -781034
[2.] Guwahati Medical College & Hospital, Guwahati, Assam, India -781006

Corresponding author(s)

1. Lipi B. Mahanta, email: lbmahanta@iasst.gov.in

Abstract

While a publicly available benchmark dataset provides a base for the development of new algorithms and comparison of results, hospital-based data collected from the real-world clinical setup is also very important in AI-based medical research for automated disease diagnosis, prediction or classification as per standard protocol. Primary data must be constantly updated so that the developed algorithms achieve as much accuracy as possible in the regional context. This dataset would support research work related to image segmentation and final classification for a complete decision support system. Liquid-based cytology (LBC) is one of the cervical screening tests. The repository consists of total of 963 LBC images sub-divided into four sets representing the four classes: NILM, LSIL, HSIL and SCC. It comprises pre-cancerous and cancerous lesions related to cervical cancer as per standards under The Bethesda System (TBS). The images were captured in 40x magnification using Leica ICC50 HD microscope collected with due consent from 460 patients visiting the O&G department of the public hospital with various gynaecological problems. The images were then viewed and categorised by experts of the pathology department.

Keywords

Cervical cancer; pap smear; liquid-based cytology; 40x; cervical pre-cancerous lesions; cervical cancerous lesions

Specifications Table

Subject	Computer Science, Computer Vision and Pattern Recognition,
Specific subject area	Medical Image Processing, Cervical Cancer, Cell segmentation, Cell classification
Type of data	Images
How data were acquired	Images were captured using a Leica DM 750 microscope with camera model ICC50 HD, in 400x (40x objective lens× 10x eyepiece) magnifications (size 2048 × 1536 pixels).
Data format	Raw JPG
Parameters for data collection	Images were captured in 400x (40x objective lens× 10x eyepiece) magnifications. The size of the images is 2048 × 1536 pixels.
Description of data collection	Liquid-based cytology provides more uniform fixation with a cleaner background and well-preserved samples for further HPV tests other than conventional Pap tests and hence it is preferred here. The LBC pap smear slides were collected from three distinguished medical diagnostic centers of the NER regions, India namely Babina Diagnostic Pvt. Ltd, Imphal, Gauhati Medical College and Hospital, Guwahati and Dr B. Barooah Cancer Institute, Guwahati. All samples involve ethical clearance protocol from the three diagnostic centers along with patient consent from total of 460 patients undergoing cervical screening tests. The images were captured in 400x magnifications using Leica DM 750 microscope, model ICC50 HD connected with the camera and a high-configured computer and software. The images represent the sub-categories of cervical lesions (malignant and pre-malignant) as NILM (Negative for Intraepithelial lesions), LSIL (Low-grade intraepithelial lesions), HSIL (High-grade intraepithelial lesions), and SCC (Squamous Cell Carcinoma).
Data source location	1. Babina Diagnostic Pvt. Ltd, Imphal, India 2. Dr B. Borooah Cancer Research Institute, Guwahati, Assam, India 3. Gauhati Medical College and Hospital, Guwahati, Assam, India
Data accessibility	Hussain, Elima; (2019), “Liquid-based cytology pap smear images for multi-class diagnosis of pre-cancerous and cancer lesions related to cervical cancer”, Mendeley Data, V4, doi: 10.17632/zddtpgzv63.4
Related research article	Co-submitted in Tissue and Cell entitled ”A comprehensive study on the multi-class diagnosis of Pap smear images using a fusion-based decision from ensemble deep convolutional neural network”

Value of the Data

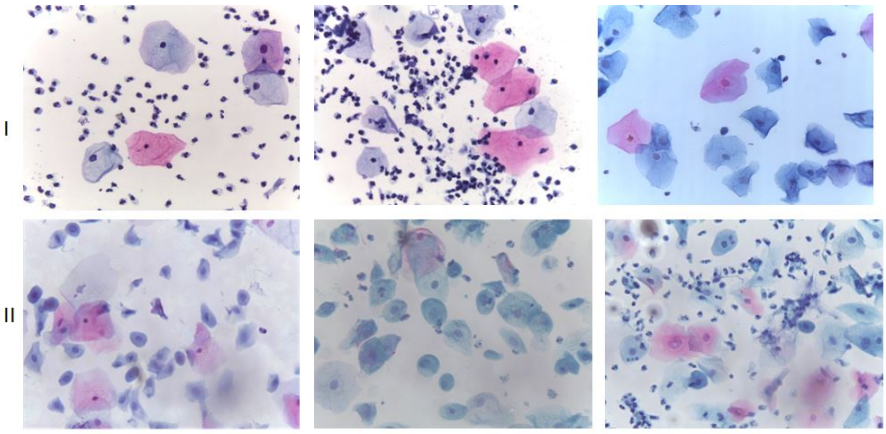
- This dataset can be used for a comparative assessment of one's experimental findings against publicly available benchmark conventional (Bora et al., 2017; Jantzen and Dounias, 2006; Lu et al., 2015; M. E. Plissiti, P. Dimitrakopoulos, G. Sfikas, C. Nikou, O. Krikoni, 2018) and Thin-Prep LBC datasets (Phoulady and Moutan, 2018).
- Researchers can use this dataset for computer-assisted diagnosis of cervical cancer which enables interpretation of such images for different image segmentation algorithms, feature extraction or feature selection methodologies and in final classification step (both binary as well as multi-class classification). In the case of binary classification (normal vs. abnormal class), the NILM category can be grouped as normal whereas LSIL, HSIL, and SCC further grouped as abnormal classes.
- Deep learning methodologies concerned with classification or semantic segmentation tasks can also be incorporated with further data augmentation techniques using these images.

Data

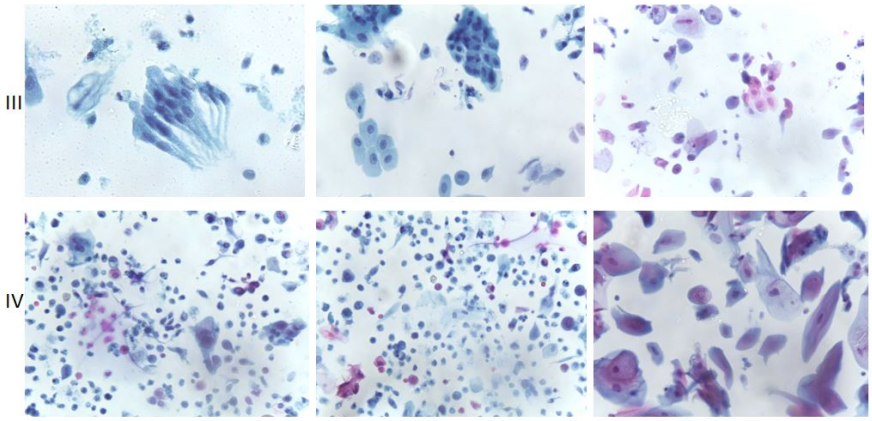
The dataset has been sub-divided into four categories each depicting the four classes of cervical cancer as per TBS standards. **Table 1.** quantifies the total images belonging to each category, a few samples of which are illustrated in **Figure 1**. The images can be used for binary and multi-class classification tasks using machine learning as well as deep learning approaches. The classification step can be integrated with image pre-processing, image segmentation and feature extraction steps which require quantitative analysis of detection of abnormal features based on cell-level morphometry like shape, color or texture analysis. This will enable automated or computer-assisted diagnosis for early detection of pre-cancerous lesions to combat cervical cancer disease. This will contribute to rapid prognosis therapy.

Table 1: Dataset arrangement

Category	No. of images
NILM	613
LSIL	163
HSIL	113
SCC	74
Total images	**963**



(A)



(B)

Figure. 1: (A) Images belonging class NILM and LSIL and **(B)** Images belonging to class HSIL and SCC

Experimental Design, Materials, and Methods

Images in the datasets were collected using liquid-based cytology (LBC) (sure-path) technique in the Obstetric and Gynecology department of Gauhati Medical College and Hospital, the primary public healthcare center of the region. LBC technique basically involves a small brush to collect the sample with target from transformation zone (where columnar epithelial cell undergoes changes into squamous epithelial cell) in the same way as a conventional smear test, but instead of transferring the smear specifically to a microscopic slide, the samples are kept into a container with additive fluid. This fluid deals with evacuating different types of unwanted debris, like mucus, blood cells, etc., prior to setting a layer of cells on the slides. The vial containing cervical samples was finally placed at a vortex with 3000 rpm for 15-20 seconds in order to break mucotic and blood particles. After adding density reagent to the sample, it undergoes sedimentation and centrifugation at 2500 rpm for 5 minutes. This is mainly done so that particles having heavy molecular weight get settled down at the bottom of the slide. After one or two alcohol wash, the slides were stained using Haemotoxylin and Eosin (H&E) staining protocol.

These slides were then used to capture images using Leica ICC50 HD microscope at 400x. 400x magnification provides better view of smear level image per slides than 100x and 200x with distinct cellular features as per the concerned categories. Ten best quality images per slides were acquired and maintained in a simple excel file along with medical reports per patient. While capturing these images, it is ensured that minimal overlap of image sections in a particular slide is happening. So images were essentially acquired by moving the microscope eyepiece over the slides in a sequential pattern. Although there is a probability of subjective error in this process, this sequence is repeated throughout to keep this error at a minimal percentage. The images were categorized as NILM, LSIL, HSIL and SCC based on the patient's report and finally confirmed with an expert pathologist’s review from pathology department.

These images may now undergo different image processing tasks subjective to computer vision and machine learning fields.

Acknowledgments

We acknowledge the Department of Biotechnology (DBT), Govt. of India for providing funds (grant no-DBTNER/Health/48/2016). Authors would like to thank the Department of Obstetrics & Gynaecology, Guwahati Medical College & Hospital, Bhangagarh, Guwahati, Assam (GMCH) for LBC set up. Authors would also like to acknowledge Dr. Anup K. Das, Senior Pathologists, Arya Wellness Centre, Guwahati, Assam for his valuable guidance mostly during the data acquisition phase and throughout. Lastly, authors would also like to thanks Dr. Dhabali Singh, Senior Pathologists, Babina Diagnostics, Imphal for contributing adequate LSIL, HSIL and SCC slides.

Transparency document

Transparency documents associated with this article can be found in the online version at http://dx.doi.org/10.17632/zddtpgzv63.4

References:

Bora, K., Chowdhury, M., Mahanta, L.B., Kundu, M.K., Das, A.K., 2017. Automated classification of Pap smear images to detect cervical dysplasia. Comput. Methods Programs Biomed. 138, 31–47. https://figshare.com/articles/Pap_smear_classification/3635715/1

Jantzen, J., Dounias, G., 2006. The Pap Smear Benchmark, in: Proceeding of NISIS-2006 Symposium. http://labs.fme.aegean.gr/decision/downloads.

Lu, Z., Carneiro, G., Bradley, A.P., 2015. An improved joint optimization of multiple level set functions for the segmentation of overlapping cervical cells. IEEE Trans. Image Process. 24, 1261–1272. https://doi.org/10.1109/TIP.2015.2389619

M. E. Plissiti, P. Dimitrakopoulos, G. Sfikas, C. Nikou, O. Krikoni, A.C., 2018. SIPAKMED: A new dataset for feature and image based classification of normal and pathological cervical cells in Pap smear images, in: IEEE International Conference on Image Processing (ICIP) 2018, Athens, Greece, 7-10 October 2018. http://www.cs.uoi.gr/~marina/sipakmed.html

Phoulady, H.A., Moutan, P.R., 2018. A New Cervical Cytology Dataset for Nucleus Detection and Image Classification (Cervix93) and Methods for Cervical Nucleus Detection, in: CVPR. https://github.com/parham-ap/cytology_dataset