Automated Diagnosis of Cervical Intraepithelial Neoplasia in Histology Images via Deep Learning

Bum-Joo Cho
Hallym University Sacred Heart Hospital, Hallym University College of Medicine
Jeong Won Kim (jwkim@hallym.or.kr)
Kangnam Sacred Heart Hospital
Jungkap Park
Hallym University Medical Center
Gui Young Kwon
Seoul Clinical laboratories
Mineui Hong
Chung-Ang University Hospital, Chung-Ang University College of Medicine
Si-Hyong Jang
Soonchunhyang University Cheonan Hospital, Soonchunhyang University College of Medicine
Heejin Bang
Konkuk University Medical Center, Konkuk University School of Medicine
Gilhyang Kim
Kangnam Sacred Heart Hospital
Sung Taek Park
Kangnam Sacred Heart Hospital

Research Article

Keywords: cervical intraepithelial neoplasia, histology image, artificial intelligence, deep learning, convolutional neural network

Posted Date: September 9th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-877842/v1

Version of Record: A version of this preprint was published at Diagnostics on February 21st, 2022. See the published version at https://doi.org/10.3390/diagnostics12020548.

Abstract

Artificial intelligence has enabled the automated diagnosis of several cancer types. We aimed to develop and validate deep learning models that automatically classify cervical intraepithelial neoplasia (CIN) based on histological images. Microscopic images of CIN3, CIN2, CIN1, and non-neoplasm were obtained. The performances of two pre-trained convolutional neural network (CNN) models adopting DenseNet-161 and EfficientNet-B7 architectures were evaluated and compared with those of pathologists. The dataset comprised 1,106 images from 588 patients; images of 10% of patients were included in the test dataset. The mean accuracies for the 4-class classification were 88.5% (95% confidence interval [CI], 86.3–90.6%) by DenseNet-161 and 89.5% (95% CI, 83.3–95.7%) by EfficientNet-B7, which were similar to human performance (93.2% and 89.7%). The mean per-class area under the receiver operating characteristic curve values by EfficientNet-B7 were 0.996, 0.990, and 0.971 in the non-neoplasm, CIN3, CIN1 groups, respectively. The class activation map detected the diagnostic area for CIN lesions. In the 3-class classification of CIN2 and CIN3 as one group, the mean accuracies of DenseNet-161 and EfficientNet-B7 increased to 91.4% (95% CI, 88.8–94.0%) and 92.6% (95% CI, 90.4–94.9%), respectively. CNN-based deep learning is a promising tool for diagnosing CIN lesions on digital histological images.

Introduction

In 2018, cervical cancer ranked as the fourth most frequently diagnosed cancer and the fourth leading cause of cancer-related death in women worldwide [1]. Despite the decreasing incidence in developed countries due to active screening and vaccination for human papilloma virus (HPV), its prevalence and mortality are increasing in sub-Saharan Africa, southeastern Asia, eastern Europe, and South America. Histologically, the most common type is squamous cell carcinoma, and HPV is the virtually necessary (but not sufficient) cause of cervical cancer [1]. For early detection, screening methods, such as the HPV test, cervical cytology, and colposcopy, are recommended; however, the gold standard for diagnosing cervical lesions is the microscopic evaluation of histopathology by a qualified pathologist [2].

Premalignant lesions of the cervix, cervical intraepithelial lesions (CINs) are traditionally graded as CIN1, CIN2, and CIN3, according to the extent of abnormal proliferation in the basal layer with increased nuclear:cytoplasmic (N:C) ratio and mitotic activity [3]; in CIN1, encompassing condyloma and mild dysplasia, atypical proliferation and mitosis occur up to the lower third of the epithelium along with koilocytotic atypia with clearly retained features of maturation. CIN2, including moderate dysplasia, demonstrates atypical basaloid cells and mitotic activity extending into the lower two-thirds of the epithelium, but with maturation in the uppermost cell layers. CIN3, encompassing severe dysplasia and carcinoma in situ, shows full-thickness atypia and mitotic activity without maturation in the top-most epithelial layers. Approximately 60% of the CIN1 lesions regress without treatment, and less than 1% progress to invasive cancer. Meanwhile, 5% of the CIN2 and 12% of CIN3 cases progress to invasive cancer if left untreated. Therefore, the clinical management of CINs entirely depends on the histologic diagnosis [2]. However, pathologists often encounter difficulties in accurately diagnosing and grading CIN [4]. The effects of inflammation, repair, pregnancy, and atrophy, as well as the inherent difficulty in

distinguishing lesions with a morphologic spectrum, complicate it and may lead to substantial inter-observer and intra-observer variability [4–6]. The time pressure, workload, and limited experience of the pathologist may be other hindrances. With the increase in cervix specimens due to population growth, increased prevalence of cancers, and longer life spans, these obstacles will likely worsen in the future. Due to the limited well-trained pathology workforce, the quality of pathology services is uneven nationwide and worldwide [7]. The use of automatic histology image classification can alleviate the scarcity in professional resources and heavy workloads.

With the advancement of artificial intelligence (AI), machine learning techniques can be used as a major ancillary tool for diagnosing tumors in various organs based on the histological images. However, most recent studies have applied techniques for the detection and classification of invasive cancers [8–13] rather than intraepithelial or premalignant lesions. With regard to cervical lesions, some studies have been devoted to the creation of computer-assisted reading systems for assessing cervical cytology specimens [14], and only a limited number of studies have focused on examining CINs [15–21]. In this study, we aimed to develop and assess an optimal convolutional neural network (CNN) model for classification of CINs.

Results

A total of 1,106 images from 588 patients were included in this study. The patients' mean age was 43.0 ± 12.4 years (range: 16–84 years). The patients' data are presented in Table 1. Non-neoplastic lesions comprised the majority class (343 cases from 250 patients, 31.0%) in the whole dataset, while CIN2 was the least common type (231 cases, 20.9%). The test dataset for the human performance evaluation comprised 117 images from 68 patients.

Table 1
Data composition for the first splitting of the training and test datasets

	Whole Dataset		Training set		Test set	
	Image N	Patient N	Image N	Patient N	Image N	Patient N
Overall	1106	588	989	542	117	68
CIN 3	266	183	236	165	30	19
CIN 2	231	108	210	97	21	11
CIN 1	266	143	234	129	32	14
Non-neoplasm	343	250	309	225	34	25

N, numbers; CIN, cervical intraepithelial neoplasia

Four-class classification performance of deep learning models and human pathologists

The mean accuracies for the 4-class classification (CIN3, CIN2, CIN1, and non-neoplasm) in the test dataset were 88.5% (95% confidence interval [CI], 86.3–90.6%) by DenseNet-161 and 89.5% (95% CI, 83.3–95.7%) by EfficientNet-B7, respectively (Supplementary Table). The validation accuracy reached a plateau within 20 epochs during the model training, as shown in Fig. 1. The overall accuracies for the 4-class classification of human pathologists were 93.2% and 89.7%, respectively. The heatmaps for the confusion matrix of the best-performing models for the test dataset and human pathologists are presented in Fig. 2.

The per-class performances of the deep learning models are presented in Table 2 and Fig. 3 depicts the per-class receiver operating characteristic (ROC) curves for the best-performing CNN models. For both CNN architectures, the mean area under the ROC curve (AUC) was highest in discriminating non-neoplastic lesions (0.996 for DenseNet-161 and 0.996 for EfficientNet-B7). For both CNN architectures, the mean AUC was lowest in discriminating CIN2 lesions, but the individual AUCs remained high (0.947 for DenseNet-161 and 0.956 for EfficientNet-B7, respectively). In determining CIN3 lesions, EfficientNet-B7 showed a mean sensitivity of 97.5% (95.4–99.5%) and a mean specificity of 96.3% (94.1–98.6%). For the CIN1 lesions, EfficientNet-B7 presented a mean sensitivity of 85.2% (73.3%-97.1%) and a mean specificity of 96.3% (95.1–97.6%).

Table 2
Per-class performances of the deep learning models in the 4-class classification

Model / Class	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	AUC (95% CI)
DenseNet-161					
CIN3	95.3 (93.7–96.8)	94.4 (93.1–95.6)	85.0 (82.3–87.7)	98.3 (97.8–98.9)	0.989 (0.982–0.996)
CIN2	75.2 (67.7–82.8)	94.1 (91.8–96.4)	76.1 (62.3–89.9)	93.8 (92.5–95.0)	0.947 (0.932–0.963)
CIN1	82.1 (77.8–86.5)	98.3 (97.4–99.2)	94.2 (92.2–96.2)	94.5 (92.6–96.4)	0.979 (0.968–0.990)
Non-neoplasm	95.6 (90.9–100.0)	98.0 (96.3–99.7)	95.0 (91.0–99.0)	98.3 (96.6–100.0)	0.996 (0.991-1.000)
EfficientNet-B7					
CIN3	97.5 (95.4–99.5)	96.3 (94.1–98.6)	90.0 (84.2–95.8)	99.1 (98.4–99.8)	0.990 (0.981–0.999)
CIN2	73.0 (62.2–83.9)	96.7 (93.7–99.7)	86.8 (75.2–98.4)	93.6 (92.3–94.8)	0.956 (0.946–0.967)
CIN1	85.2 (73.3–97.1)	96.3 (95.1–97.6)	88.5 (88.2–88.8)	95.5 (91.3–99.8)	0.971 (0.950–0.993)
Non-neoplasm	95.6 (90.9–100.0)	96.3 (92.2–100.0)	92.3 (84.8–99.8)	98.3 (96.6–100.0)	0.996 (0.992–0.999)

PPV, positive predictive value; NPV, negative predictive value; AUC, area under the receiver operating characteristic curve; CI, confidence interval

Histologic review of misclassified cases in four-class classification using best-performing CNN models

No false-positive cases were included in the best-performing CNN models, both DenseNet-161 and EfficientNet-B7 (Fig. 2). False-negative cases were not observed by EfficientNet-B7; among 117 test cases, three CIN1s (2.6%) were classified as false-negative cases by DenseNet-161. After a histological review, it appeared that the scarcity of characteristic koilocytotic cells might have contributed to the misclassification (Supplementary Fig. a).

Eight (9.2%) out of 87 CIN cases were unsuccessfully graded by DenseNet-161; one CIN3 (1.1%) and two CIN2 (2.3%) cases were downgraded as CIN1 and CIN2, respectively, while one CIN1 (1.1%) and four CIN2 cases (4.6%) were upgraded as CIN2 and CIN3, respectively (Fig. 2a). EfficientNet-B7 misgraded the five CIN2 cases (5.7%): four cases were classified as CIN1, while one case was classified as CIN3 (Fig. 2b). None of the CIN1 cases were classified as CIN3, and none of the CIN3 cases were classified as CIN1. On

histological review, histology of CIN3 cases downgraded as CIN2 was not sufficient to be classified as carcinoma in situ. The cases showed basal/parabasal-type atypia throughout the full-thickness of the epithelium (Supplementary Fig. b). CIN2 cases downgraded as CIN1 had atypia extending to the lower half of the epithelium with koilocytotic changes in the upper half and maturation in the uppermost layers (Supplementary Fig. c). The CIN1 case upgraded as CIN2 showed disoriented epithelium (Supplementary Fig. d). One of the CIN2 cases upgraded as CIN3 showed atrophy (Supplementary Fig. f).

Three-class classification performance of deep learning models and human pathologists

In the 3-class classification discriminating the images into CIN2–3, CIN1, and non-neoplasm, the mean accuracies in the test dataset increased up to 91.4% (95% CI, 88.8–94.0%) by DenseNet-161 and 92.6% (95% CI, 90.4–94.9%) by EfficientNet-B7. The overall accuracies for the 3-class classification of human pathologists were 95.7% and 92.3%, respectively. Figure 4 shows the heatmaps of the confusion matrix of the best-performing models for the test dataset and human pathologists.

The per-class performances of the deep learning models in the 3-class classification are listed in Table 3. The mean AUCs for non-neoplastic lesions were 0.996 (95% CI, 0.992–0.999) for DenseNet-161 and 0.993 (95% CI, 0.985–1.000) for EfficientNet-B7. The mean AUCs for CIN2–3 and CIN1 were 0.981 and 0.974 for DenseNet-161 and 0.982 and 0.979 for EfficientNet-B7. In terms of determining CIN2–3 lesions, EfficientNet-B7 showed a mean sensitivity of 94.8% (92.7–96.7%) and a mean specificity of 93.4% (90.1–96.8%).

Table 3
Per-class performances of the deep learning models in the 3-class classification

Model / Class	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	AUC (95% CI)
DenseNet-161					
CIN2-3	92.0 (86.9–97.1)	92.4 (85.3–99.6)	92.5 (87.0–98.0)	93.4 (90.2–96.7)	0.981 (0.973–0.989)
CIN1	80.9 (70.9–90.8)	96.0 (94.2–97.7)	87.0 (84.0-89.9)	94.5 (93.3–95.6)	0.974 (0.968–0.980)
Non-neoplasm	97.8 (94.2–100.0)	97.5 (95.6–99.5)	94.4 (90.0-98.9)	99.1 (97.6–100.0)	0.996 (0.992–0.999)
EfficientNet-B7					
CIN2-3	94.8 (92.8–96.7)	93.4 (90.1–96.8)	92.9 (90.3–95.6)	95.1 (92.3–97.9)	0.982 (0.971–0.993)
CIN1	86.1 (82.4–89.7)	96.4 (95.2–97.5)	87.6 (81.2–94.0)	95.6 (94.3–96.9)	0.979 (0.972–0.985)
Non-neoplasm	94.7 (92.8–96.6)	98.4 (97.0-99.7)	96.0 (92.8–99.2)	97.8 (97.1–98.6)	0.993 (0.985-1.000)

PPV, positive predictive value; NPV, negative predictive value; AUC, area under the receiver operating characteristic curve; CI, confidence interval

Analysis of Grad-CAM images by CNN model

Figure 5 shows the representative Grad-CAM images of non-neoplasms, CIN1, CIN2, and CIN3. Grad-CAM images were reviewed by a pathologist, and the region of interest of the deep learning model agreed with that of humans. The CNN model successfully detected squamous epithelium and recognized images from the transformation zone and exocervix, atrophic cervix, and cervicitis with erosion as non-neoplasms. In Grad-CAM images, CIN1, CIN2, and CIN3 characterized by presence of koilocytotic cells or hyperchromatic atypical cells with a high nuclear/cytoplasmic ratio and increased mitotic activity were depicted as highlighted areas. According to the distribution of abnormal cells, different layers of squamous epithelium were highlighted.

Discussion

In recent years, AI has been used in the field of pathologic image diagnosis, and many studies have shown promising results in detecting and diagnosing cancers in a variety of organs, including the stomach, breast, skin, prostate, brain, and lung [13, 22–24]. As for cervical cancer, with the advancement in the management of preinvasive lesions, the increasing diagnostic workload of cervical biopsy calls for the development of high-performance algorithms with high sensitivity and specificity. Practically, many pathologists experienced more difficulty and burden in accurately classifying preinvasive CIN lesions

than in distinguishing between invasive and non-neoplastic condition [4]. Therefore, we focused on developing an optimized CNN system for CIN grading.

Two CNN architectures, DenseNet-161 and EfficientNet-B7, were adopted in our study. EfficientNet-B7 is a recently developed heavy model and a state-of-the-art architecture; it showed better performance than DenseNet-161 [25]. However, DenseNet-161 also showed excellent performance and presented good cost-effectiveness [26]. Through repeated validation and tests of two CNN models, we determined the optimal image preprocessing conditions (640 × 480 pixels size and normalization of each RGB color channel in the ImageNet dataset) in which the CNN models can achieve better performance. In addition, we found that data augmentation and histogram equalization did not improve the model performance. In the 4-class classification, the mean accuracies for DenseNet161 and EfficientNet-B were 88.5% and 89.5%, respectively, and the performance was similar to that of human pathologists (93.2% and 89.7%, respectively). The mean AUC values of both CNN models were considerably high in all four classes (Table 1). Furthermore, the mean accuracies of both models for 3-class classification were increased to 91.4% and 92.6% by DenseNet-161 and EfficientNet-B7, respectively, which are almost the same levels as those of human pathologists (95.7% and 92.3%, respectively).

In the natural clinical course, CIN1 has a low potential for progression and a high potential for regression, whereas CIN2 and CIN3 have a higher potential for progression and a lower potential for regression [3]. CIN3 is a direct precursor of invasive cervical cancer, and active treatment is recommended. By contrast, CIN1 observation is the preferred approach. Therefore, it is much more critical to quickly determine CIN1 or CIN3 than CIN2 or CIN3. Despite some misclassification of CIN2, EfficientNet-B7 perfectly discriminated CIN1 from CIN3 based on the 4-class classification (Fig. 2), and the results demonstrated its clinical applicability.

We observed that the CNN models have a weakness in classifying CIN2 (75.2% and 73.0% sensitivities for DenseNet-161 and EfficientNet-B7, respectively); considering that it is often challenging for pathologists to distinguish CIN2 from CIN1 and CIN3 and inter-observer agreement is notoriously poor at this interface, even among experts [4], the performance of these CNN models is almost similar to that of human pathologists even in this respect. Moreover, the difficulty in classifying CIN2 can be attributed to its inherent nature, which is intermediate in the morphological spectrum of CIN. Due to the ambiguity of CIN2 diagnosis based on the hematoxylin and eosin (H&E) morphology, the LAST Project suggested that the addition of p16 immunohistochemical stain significantly improves the reliability of CIN2 diagnosis and advised the use of p16 staining to confirm the presence of a high-grade lesion when CIN2 is diagnosed based on H&E slide [2]. In future studies, analyzing H&E images along with the results of p16 immunohistochemical staining would be helpful to increase diagnostic accuracy of CNN models.

For determining the CIN1 lesions, the mean sensitivities were 82.1% and 85.2% by DenseNet-161 and EfficientNet-B7, respectively, which were lower than those of CIN3 and non-neoplasms. On histologic review, the scarcity of characteristic koilocytotic cells in CIN1, severe inflammation, and metaplastic changes might have contributed to the inaccuracy of CNN classification. For more precise detection of

koilocytotic cells, the CNN model needs to be improved. To reduce the false-positive rate, more variable non-neoplastic lesions, such as chronic cervicitis, metaplastic mucosa, and atrophy, should be included in the study set, and a repeat validation would be helpful.

Automated screening machines have been developed for analyzing cervical cytology smears, and a few FDA-approved automated primary screening device are available [14]. However, it is more difficult to develop an automated tool for cervical tissue histology due to the complexity of the patterns observed and the structural associations between different tissue components [17]. Keenan et al. developed a machine vision system for histological grading of CIN using the KS400 macro programming language. It was a scoring system that analyzes geometric data, and 62.7% of the CIN cases with captured images were correctly classified [17]. Several previous studies have used multiclass support vector machines and gray-level co-occurrence matrices to analyze whole slide images (WSIs) or selected images [18, 21, 27]. Despite some promising results, the small data size of less than 100 cases with insufficiently validated or curated images and the extremely complicated methodology limited the applicability of the study results. Huang et al. proposed a method based on the least absolute shrinkage and selection operator and ensemble learning support vector machine [19]. They showed that the accuracy of normal-cancer classification was high (99.64%), but the accuracy of the low-grade squamous intraepithelial lesion (LSIL)-high-grade squamous intraepithelial lesion (HSIL) classification was 76.34%. A recent study that classified cervical tissue pathological images based on fusing deep convolution features has been published [28]. The researchers analyzed the dataset comprising small-sized images cropped from 468 WSIs, including those of normal tissues, LSIL, HSIL, and cancer; Resnet50v2 and DenseNet121(C_5) showed excellent performance, with an average classification accuracy of 95.33%.

Pathologic classification is an image-based method, and CNN is an optimized AI tool for image learning. Our study showed that CNN is a robust instrument for pathologic classification, but some things must be considered. For CNN to be developed and to work properly, collecting a large amount of accurate data is of utmost importance. Because CNN produces results very faithfully in the learned input, the quality of the CNN output absolutely depends on the quality of the input data. In order to develop a clinically relevant CNN model for pathologic diagnosis, a superb dataset from expert pathologists must be constructed. Recently, Meng et al. provided a public cervical histopathology dataset for computer-aided diagnosis, called MTCHI [20]. Pathologic diagnosis is sometimes equivocal and might be challenging to perform in some lesions in the gray zone or lesions with reactive changes. Therefore, pathologists should continue to improve and make an objective pathological diagnosis. A limitation of high-quality H&E slide images is the need for using AI to perform a pathologic diagnosis. Although staining and mounting are automated, preparing pathology slides, sectioning, and embedding are still manually performed. Artifacts in the production process, such as tissue overlapping, tangential embedding, and poor sectioning, hinder the acquisition of focused images and cause AI to make diagnostic errors.

Hence, we aimed to develop an artificial technique for classifying CIN from the WSI of cervical biopsy, but some practical difficulties were observed. In the WSIs of tissues, grading of intraepithelial neoplasia or dysplasia is much more complicated than finding lesions or cancer. Because CIN is a morphological

spectrum, cervical biopsy specimens show large differences in disease degrees and mix of lesions. This makes it difficult for pathologists to precisely annotate according to the CIN grade in small biopsies. Compared with other tissues such as the breast, colon, and stomach, the specimen used for cervical biopsy are tissue strips or appear irregular in shape and often include a small amount of epithelium. In addition, it is easily embedded in a disoriented or tangential manner. These were obstacles in making a standardized dataset using WSIs suitable for training and validation of the CNN model. In this study, we built a reliable dataset of CIN provided by three qualified pathologists and analyzed the CNN performance prior to its application in WSI. The dataset and the advanced CNN model, EfficientNet-B7, might be applied in future research using the WSI of cervical biopsy.

In summary, we built a reliable dataset for CIN classification and showed that EfficientNet-B7 and DenseNet-161 provided a promising performance in classifying cervical lesions on digital histology images. In terms of accuracy, EfficientNet-B7 had a functional advantage over DenseNet-161. Grad-CAM images used in the CNN models located the areas where CIN lesions can be found. Moreover, we realized that the accurate identification and classification of CIN by CNN relies entirely on the standardized diagnosis of pathologists, and the professional knowledge and analytical experience of pathologists are the cornerstone of technical advancement. An exquisite AI tool trained using a well-established and standardized dataset would be helpful in improving the pathology services worldwide.

Methods

Data collection

Female patients who were scheduled for colposcopic biopsy or conization due to suspicion of CIN at Kangnam Sacred Heart Hospital between 2015 and 2017 were retrospectively enrolled. One experienced pathologist (J.W.K.) reviewed the histological slides of tissue sections of the involved patients that were stained with H&E and obtained digital microscopic photographs of representative lesions at an objective magnification of 20× using a microscope (Olympus BX51; Melville, NY, USA) equipped with a digital camera (Olympus DP2). Photographic images were acquired in JPEG format with a resolution of 2560 × 1920 or 1280 × 960 pixels. Unsuitable blurred or defocused images were excluded from this study. All digital images were classified into four classes according to the previously suggested criteria: CIN3, CIN2, CIN1, and non-neoplasm [2]. The protocol for this study was approved by the Institutional Review Board (IRB) of Kangnam Sacred Heart Hospital (IRB no. 2018-03-13). Informed consent from the patients was waived by the Institutional Review Board of Kangnam Sacred Heart Hospital and the informed consent was obtained from two pathologists who participated for human performance evaluation. All experiments were performed in accordance with the Declaration of Helsinki.

Dataset construction

Digitalized images were re-reviewed and classified independently by two experienced pathologists (M. H. and G. Y. K.) blinded to the pre-classified results. Only images that were accurately classified by the three

pathologists were involved in this study. Ultimately, 1,106 microscopic images from 588 patients were included: 266, CIN3; 231, CIN2; 266, CIN1; and 343, non-neoplasms (Table 1).

From the whole dataset, the test dataset was randomly split three times with a ratio of 10% to evaluate the performance of the trained CNN models. Random train/test set splitting was performed for each class using the patient ID as the key to avoid the simultaneous involvement of the same class image of one patient in both the training and test datasets. For each three splitting, the training set consisted of 90% of the whole dataset and was divided into the training dataset proper and the tuning (or validation) dataset, with a ratio of 80%:10%.

Data preprocessing

All images were resized to 640 × 480 pixels and were normalized for each RGB color channel based on the mean and standard deviation values of the images in the ImageNet dataset. Data augmentation and histogram equalization were not performed as these methods did not improve the model performance in our pilot studies.

Deep learning model training

Two CNN architectures were adopted: DenseNet-161 and EfficientNet-B7. The details of the CNN models are described in previous studies [25, 26]. Briefly, DenseNet-161 is characterized by a dense block that uses the feature maps of the previous layers as the input of the current layer [26]. EfficientNet is characterized by an MBconv block that balances the width and depth of the CNN via reinforcement learning [25]. These models were pre-trained using the ImageNet Large Scale Visual Recognition Challenge dataset and fine-tuned using the training dataset of this study.

In the first experiment, the CNN models were trained to perform 4-class classification that classified images into CIN3, CIN2, CIN1, and non-neoplastic lesions. Then, the CIN2 and CIN3 groups of the whole dataset were merged into one group, representing HSIL. In the second experiment, the training/test set splitting was re-performed; the CNN models were trained to perform 3-class classification and classified the images into CIN2–3 and CIN1, which represent LSILs, and non-neoplasms.

The model was trained using the PyTorch platform with categorical cross-entropy as the loss function. The Adam optimizer was adopted with a β1 value of 0.9 and a β2 value of 0.999. The learning rate was 1e-4, and the batch sizes were 15 and 5 for DenseNet-161 and EfficientNet-B7, respectively. The number of epochs was set to 100. The hardware platform was equipped with NVIDIA GeForce GTX 1080ti 6-way graphics processing units, dual Xeon central processing units, 128 GB RAM, and a customized water-cooling system.

Saliency maps were produced to identify the regions of interest. A gradient-weighted class activation mapping (Grad-CAM) was implemented [29]. For implementation, the final few layers of the CNN models were opened, and the global average pooling and softmax layers were attached.

Human performance evaluation

For the first test dataset, two other experienced human pathologists, who were blinded to the true labels, independently classified the images, and the performances were evaluated. Human performances were compared with those of the CNN models.

Main outcome measures and statistical methods

The primary outcome was the model performance for 4-class classification, while the secondary outcome was model performance for 3-class classification. The performance of the CNN model was evaluated using three different test datasets, and the performance was estimated using means and 95% CIs. The performance was evaluated using the diagnostic accuracy and AUC. For each class, per-class sensitivity, specificity, positive predictive value, and negative predictive value were also evaluated. Continuous or categorical variables are expressed as means or percentages with 95% CIs. Statistical significance was set at $P < 0.05$.

Declarations

Funding:

This research was supported by the Bio & Medical Technology Development Program of the National Research Foundation (NRF) and funded by the Korean government (MSIT) (No. NRF-2019R1G1A1011227).

Author contributions

J.W.K and B.C. conceptualized, designed the study, coordinated data collection, carried out analysis and interpretation of data and wrote the manuscript. J.P., G.Y.K. M.H., H.B., G.K., S.J., S.T.P. collected data and carried initial analyses. All the authors reviewed the final manuscript.

Competing interests

The authors declare no conflict of interest.

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

1. Bray, F. *et al.* Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*, **68**, 394–424

https://doi.org/10.3322/caac.21492 (2018).

2. Darragh, T. M. *et al.* The Lower Anogenital Squamous Terminology Standardization Project for HPV-Associated Lesions: background and consensus recommendations from the College of American Pathologists and the American Society for Colposcopy and Cervical Pathology. *Arch Pathol Lab Med*, **136**, 1266–1297 https://doi.org/10.5858/arpa.LGT200570 (2012).
3. Mills, A. M. *et al.* Squamous intraepithelial lesions of the uterine cervix In: WHO Classification of Tumors Editorial Board. Female Genital Tumors. 5th ed. p.342 – 46 (Lyon: International Agency for Research on Cancer, 2019).
4. Stoler, M. H. & Schiffman, M. Interobserver reproducibility of cervical cytologic and histologic interpretations: realistic estimates from the ASCUS-LSIL Triage Study., **285**, 1500–1505 https://doi.org/10.1001/jama.285.11.1500 (2001).
5. Castle, P. E., Stoler, M. H., Solomon, D. & Schiffman, M. The relationship of community biopsy-diagnosed cervical intraepithelial neoplasia grade 2 to the quality control pathology-reviewed diagnoses: an ALTS report. *Am J Clin Pathol*, **127**, 805–815 https://doi.org/10.1309/pt3pnc1ql2f4d2vl (2007).
6. Carreon, J. D. *et al.* CIN2 is a much less reproducible and less valid diagnosis than CIN3: results from a histological review of population-based cervical samples. *Int J Gynecol Pathol*, **26**, 441–446 https://doi.org/10.1097/pgp.0b013e31805152ab (2007).
7. Adesina, A. *et al.* Improvement of pathology in sub-Saharan Africa. *Lancet Oncol*, **14**, e152–157 https://doi.org/10.1016/s1470-2045(12)70598-3 (2013).
8. Bulten, W. *et al.* Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study. *Lancet Oncol*, https://doi.org/10.1016/s1470-2045(19)30739-9 (2020).
9. Courtiol, P. *et al.* Deep learning-based classification of mesothelioma improves prediction of patient outcome. *Nat Med*, **25**, 1519–1525 https://doi.org/10.1038/s41591-019-0583-3 (2019).
10. Halicek, M. *et al.* Head and Neck Cancer Detection in Digitized Whole-Slide Histology Using Convolutional Neural Networks. *Sci Rep*, **9**, 14043 https://doi.org/10.1038/s41598-019-50313-x (2019).
11. Ker, J., Bai, Y., Lee, H. Y., Rao, J. & Wang, L. Automated brain histology classification using machine learning. *J Clin Neurosci*, **66**, 239–245 https://doi.org/10.1016/j.jocn.2019.05.019 (2019).
12. Yoon, H. *et al.* Tumor Identification in Colorectal Histology Images Using a Convolutional Neural Network. *J Digit Imaging*, **32**, 131–140 https://doi.org/10.1007/s10278-018-0112-9 (2019).
13. Lucas, M. *et al.* Deep learning for automatic Gleason pattern classification for grade group determination of prostate biopsies. *Virchows Arch*, **475**, 77–83 https://doi.org/10.1007/s00428-019-02577-x (2019).
14. Valente, P. T. & Schantz, H. D. Cytology automation: An overview. *Laboratory Medicine*, **32**, 686–690 (2001).
15. Guo, P. *et al.* Enhancements in localized classification for uterine cervical cancer digital histology image assessment. *J Pathol Inform*, **7**, 51 https://doi.org/10.4103/2153-3539.197193 (2016).

16. Sornapudi, S. *et al.* Deep Learning Nuclei Detection in Digitized Histology Images by Superpixels. *J Pathol Inform*, **9**, 5 https://doi.org/10.4103/jpi.jpi_74_17 (2018).
17. Keenan, S. J. *et al.* An automated machine vision system for the histological grading of cervical intraepithelial neoplasia (CIN). *The Journal of pathology*, **192**, 351–362 (2000).
18. Guo, P. *et al.* Nuclei-based features for uterine cervical cancer histology image analysis with fusion-based classification. *IEEE journal of biomedical and health informatics*, **20**, 1595–1607 (2015).
19. Huang, P. *et al.* Classification of cervical biopsy images based on LASSO and EL-SVM. *IEEE Access*, **8**, 24219–24228 (2020).
20. Meng, Z., Zhao, Z., Li, B., Su, F. & Guo, L. A Cervical Histopathology Dataset for Computer Aided Diagnosis of Precancerous Lesions. *IEEE Transactions on Medical Imaging*, **40**, 1531–1541 (2021).
21. Wang, Y. *et al.* Assisted diagnosis of cervical intraepithelial neoplasia (CIN). *IEEE Journal of Selected Topics in Signal Processing*, **3**, 112–121 (2009).
22. Park, J. *et al.* A prospective validation and observer performance study of a deep learning algorithm for pathologic diagnosis of gastric tumors in endoscopic biopsies. *Clin. Cancer Res*, **27**, 719–728 (2021).
23. Han, Z. *et al.* Breast cancer multi-classification from histopathological images with structured deep learning model. *Scientific reports*, **7**, 1–10 (2017).
24. Coudray, N. *et al.* Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nat Med*, **24**, 1559–1567 https://doi.org/10.1038/s41591-018-0177-5 (2018).
25. Tan, M. & Le, Q. in *International Conference on Machine Learning.* 6105–6114 (PMLR).
26. Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. in *Proceedings of the IEEE conference on computer vision and pattern recognition.* 4700–4708.
27. Wei, L., Gan, Q. & Ji, T. Cervical cancer histology image identification method based on texture and lesion area features. *Computer Assisted Surgery*, **22**, 186–199 (2017).
28. Huang, P., Tan, X., Chen, C., Lv, X. & Li, Y. AF-SENet: classification of cancer in cervical tissue pathological images based on fusing deep convolution features. *Sensors*, **21**, 122 (2021).
29. Selvaraju, R. R. *et al.* in *Proceedings of the IEEE international conference on computer vision.* 618–626.

Figures

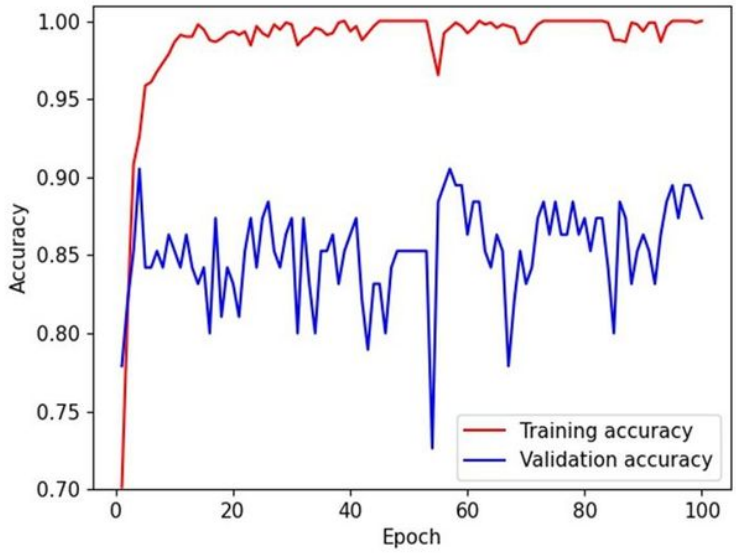

Figure 1

Figure 1

A training curve for training and validation accuracies. The validation accuracy reached a plateau within 20 epochs during model training

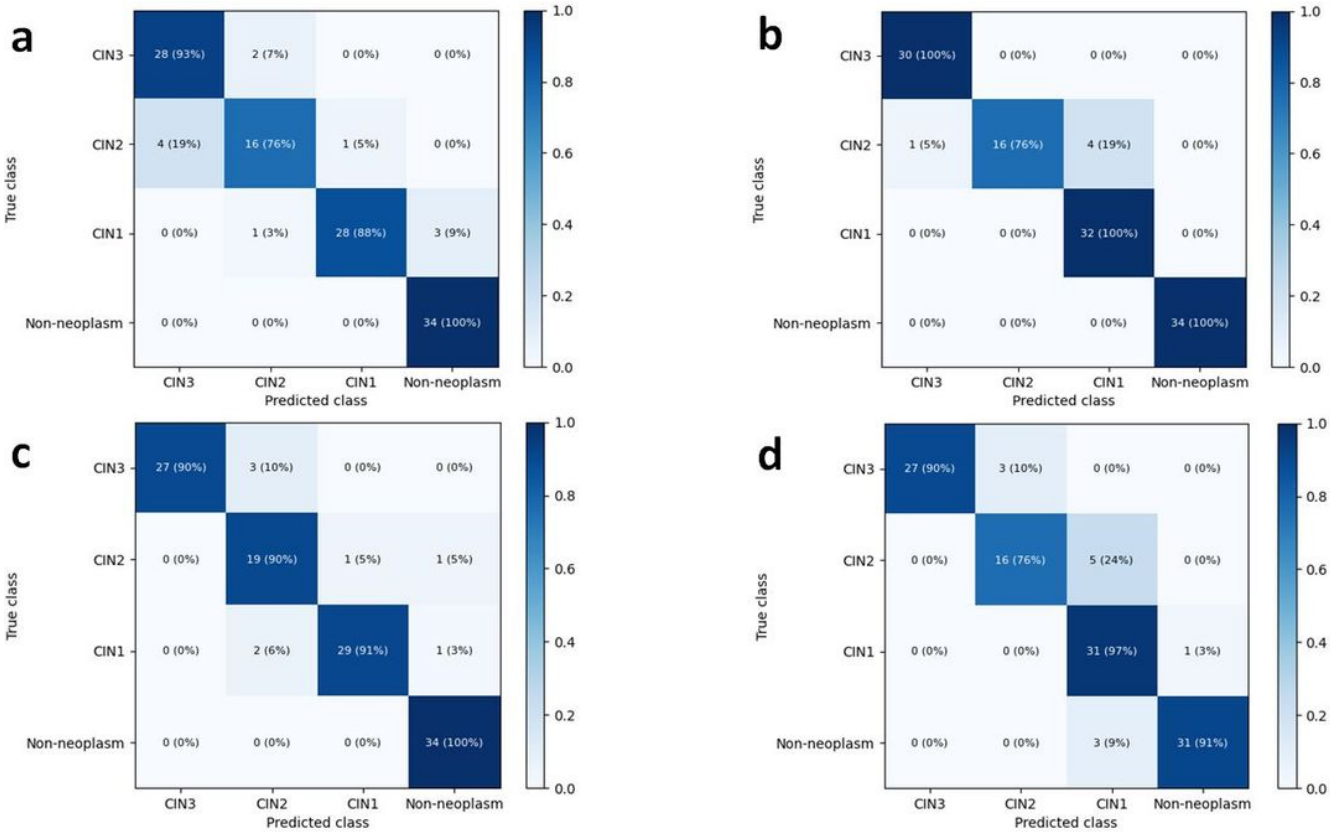

Figure 2

Figure 2

Heatmaps for confusion matrix of the best-performing CNN models and human pathologists in the 4-class classification. There were three false-negative cases in the best-performing DenseNet-161 (a) model, there was no false-negative or false-positive case with the best-performing EfficientNet-B7 (b). Pathologist 1 (c) classified CIN2 with higher sensitivity than pathologist 2 (d). CIN, cervical intraepithelial neoplasia; CNN, convolutional neural network

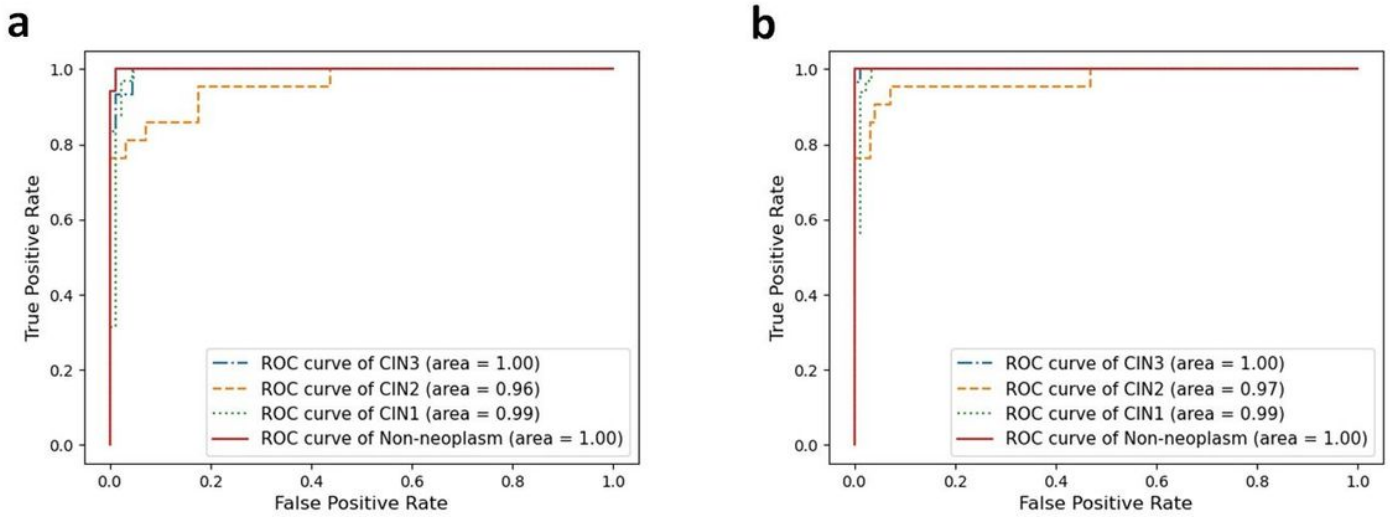

Figure 3

Figure 3

Per-class ROC curves for 4-class classification the best-performing CNN models. For DenseNet-161 (a) and EfficientNet-B7 (b) with best performance, AUC was higher in discriminating non-neoplasm and CIN3 rather than in classifying CIN2 and CIN1. AUC, area under the ROC curve; CIN, cervical intraepithelial neoplasia; CNN, convolutional neural network; ROC, receiver operating characteristic

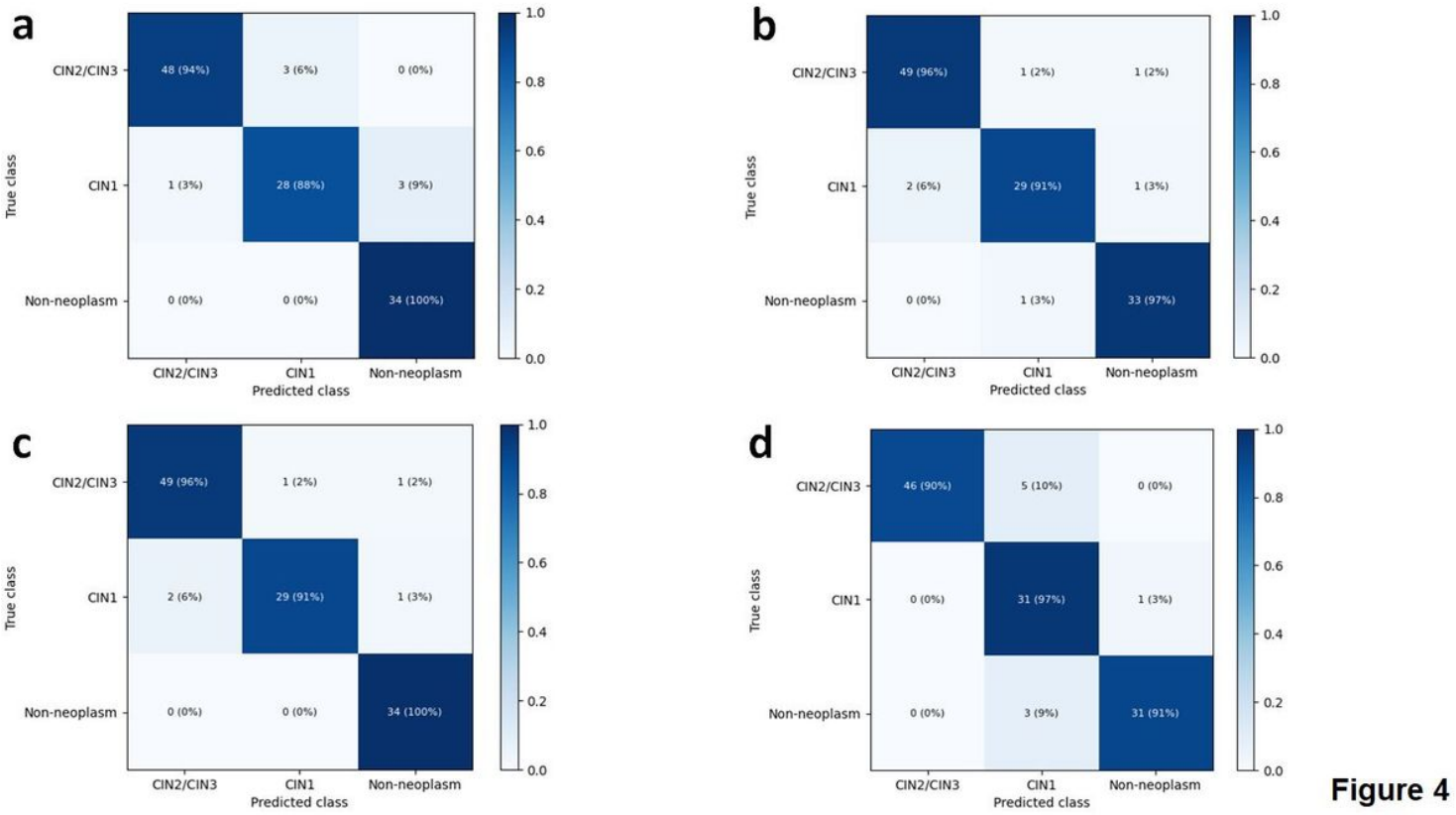

Figure 4

Heatmaps for confusion matrix of the best-performing CNN models and human pathologists in the 3-class classification. The overall accuracies increased up to 94.0% by DenseNet-161 (a) and 94.9% by EfficientNet-B7 (b), similar to those of human pathologists 1 and 2, 95.7% (c) and 92.3% (d), respectively. CNN, convolutional neural network

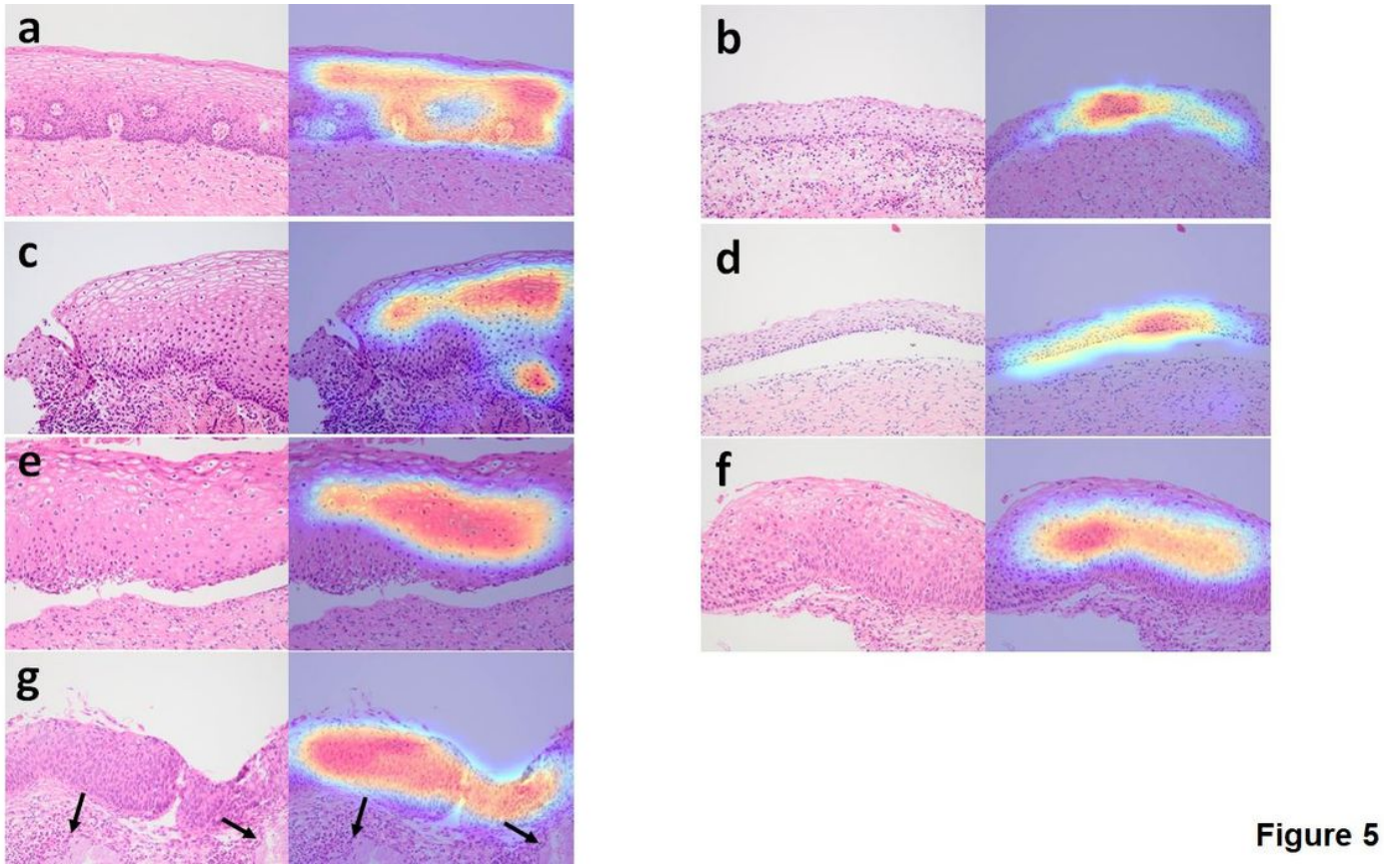

Figure 5

Figure 5

Grad-CAM images by EfficientNet-B7. Normal squamous epithelium was highlighted in Grad-CAM images (a-d). Images from cervix interpreted as non-neoplasm by the EfficientNet-B7 include exocervix (a), metaplastic mucosa from transformation zone (b), cervicitis and erosion (c) and atrophic mucosa (d). In CIN1, layers with koilocytotic cells were mainly highlighted (e). The highlighted areas extended to the upper two-third of the epithelium in CIN2 (f) and full-thickness of the epithelium in CIN3 (g). Normal endocervical glands (g, black arrows) were not highlighted. CIN, cervical intraepithelial neoplasia; Grad-CAM, Gradient-weighted class activation mapping

Supplementary Files

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

- CINCNNSRSupplefinal.docx
- SF.jpg