J Inf Process Syst, Vol.17, No.3, pp.453~461, June 2021
https://doi.org/10.3745/JIPS.04.0214

ISSN 1976-913X (Print)
ISSN 2092-805X (Electronic)

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

Attentive Transfer Learning via Self-supervised Learning for Cervical Dysplasia Diagnosis

Jinyeong Chae*, Roger Zimmermann**, Dongho Kim***, and Jihie Kim*

Abstract

Many deep learning approaches have been studied for image classification in computer vision. However, there are not enough data to generate accurate models in medical fields, and many datasets are not annotated. This study presents a new method that can use both unlabeled and labeled data. The proposed method is applied to classify cervix images into normal versus cancerous, and we demonstrate the results. First, we use a patch self-supervised learning for training the global context of the image using an unlabeled image dataset. Second, we generate a classifier model by using the transferred knowledge from self-supervised learning. We also apply attention learning to capture the local features of the image. The combined method provides better performance than state-of-the-art approaches in accuracy and sensitivity.

Keywords

Attention Learning, Cervical Dysplasia, Patch self-supervised Learning, Transfer Learning

1. Introduction

Deep learning approaches for computer vision are heavily studied. Various neural network models have been proposed for image analysis problems. For example, convolution-based models were proposed and applied to many application domains. In particular, deep learning approaches in medical areas were introduced to mitigate a shortage of data and as an effective feature extraction method from available data. To further improve the performance, we present a novel approach that uses unlabeled images and helps the model to focus on the crucial part of an image using a convolution-based model.

We apply the proposed method to a cervix image dataset to evaluate the approach. Cervical cancer is the fourth most common cancer in women. In 2018, an estimated 570,000 women were diagnosed with cervical cancer worldwide, and about 311,000 women died from the disease [1]. Cervical intraepithelial neoplasia (CIN), which represents an abnormal growth of cells which may lead to cervical cancer, can be diagnosed into three classes: CIN1 (mild), CIN2 (moderate), CIN3 (severe), according to the World Health Organization. A CIN1 infection needs constant observation, while CIN2 and CIN3 require treatment. It is necessary to separately classify normal/CIN1 from CIN2+ to detect cervical cancer.

Manuscript received February 2, 2021; first revision March 12, 2021; accepted March 26, 2021.

Corresponding Author: Jihie Kim (jihie.kim@dgu.edu)
* Dept. of Artificial Intelligence, Dongguk University, Seoul, Korea (jiny419@dgu.ac.kr, jihie.kim@dgu.edu)
** School of Computing, National University of Singapore (rogerz@comp.nus.edu.sg)
*** Dongguk Institute of Convergence Education, Dongguk University, Seoul, Korea (dongho.kim@dgu.edu)

A preliminary version of this paper was accepted at Intelligent Systems Conference (IntelliSys) 2021.

Current common cervical cancer screening methods are Pap test, human papillomavirus (HPV) test, and visual examination. Both visual inspection and invasive tests are required to achieve reasonable accuracy in diagnosing cancer. However, performing the two types of tests (visual screening and invasive tests) is challenging in many real-world environments.

In addition, an invasive test needs laboratory settings and professional medical devices. Consequently, a non-invasive visual screening is more cost-effective. For this reason, many previous studies have classified cervical cancer by using the region of interest (RoI) of a patient's cervical image. For example, in [2,3], the authors used a RoI method to detect the cervix from the cervix image, and then the RoI is used for cancer classification. In [4], they presented a LeNet model that uses the RoI pre-processing method for cervix images. Although using images for diagnosing cancer is noninvasive, a visual examination needs skilled physicians to detect the cervix shape, color, texture, and manual annotation for the RoI of the cervix images. That is why using unlabeled images is difficult in data analysis.

This study presents a new approach by using both labeled and unlabeled images to classify cancer. We offer a patch self-supervised learning to extract the global context of the unlabeled images by applying a puzzle pretext task. We also provide a new deep learning model to diagnose cervix cancer with an attention model of which the weight is transferred from self-supervised learning. By adding the transferred knowledge and attention to the deep learning model, the results also become interpretable by presenting which part of the image is used for the diagnosis. These methods neither require additional EHR data nor manual annotation of the image's RoI to improve the performance.

Our contributions are three-fold. First, we present a new patch self-supervised learning for pretraining with unlabeled cervix images. Second, we propose a deep learning model transferred from self-supervised learning for diagnosing cervical cancer. Third, we propose new attention learning for capturing RoI and extracting visual features without additional labeled data. The following section discusses related work. The description of the data used in our experiments is provided in the data section. We also discuss our method's details in the method section, while experiments and results are given in the results section. Finally, we present conclusions and future work in the last section.

2. Related Works

2.1 Self-supervised Learning

A lot of research for learning without requiring manual annotation is currently ongoing to harvest the vast amount of visual data available today. The study of Gidaris et al. [5] learned image features by training convolutional nets to recognize rotations applied to images. In the PASCAL VOC 2007 detection task, their unsupervised pre-trained AlexNet model achieved a state-of-the-art (among unsupervised methods) mAP of 54.4%, which was only 2.4 points lower than that of the supervised case. The study presented a rotation classification for a pretext task. However, since the cervix images in our study are not affected by rotation, the task is not appropriate. Chen et al. [6] proposed data augmentation for contrastive self-supervised learning algorithms without requiring specialized architectures, a memory bank, and a linear classifier. The study achieved a 76.5% top-1 accuracy, a 7% relative improvement over the previous state-of-the-art, matching the performance of a supervised ResNet50. It studied 10 data

augmentation operators for the pretext task, but the data augmentation required a lot of memory. In [7], the authors presented Momentum Contrast (MoCo) for unsupervised visual representation learning. Contrastive learning as a dictionary look-up builds a dynamic dictionary with a queue and a moving-averaged encoder. In detection/segmentation tasks on PASCAL VOC and COCO, MoCo can be superior to its supervised pretraining counter-part. They also achieved better memory efficiency than [6], but the study required a large dictionary which is affected by datasets. For training a global context, we applied puzzle self-supervised learning for predicting the relative position of the image using the unlabeled image. We considered only 1,000 combinations out of the 9! cases and used a weight-sharing model for efficient memory usage.

2.2 Cervical Cancer Analysis

Several deep learning approaches using cervix images have been proposed for cervix cancer analyses. In [3,8], the authors used the RoI method to detect the cervix in the image. They applied a CNN model to classify cervix type and cervix cancer, but this method needs manual annotations with skilled expertise, and the labeled dataset is not publicly available. In [9], the authors manually cropped images and applied data augmentation using rotation, horizontal flip, and vertical flip to generate a CNN model to classify cervix cancer. This method could be biased when the annotations are unreliable. Our study presents a new self-supervised learning using unlabeled images and proposes an attention learning to focus on the critical part of the image. In [4,10], the authors showed convolution-based models with LeNet and AlexNet for classifying cervical cancer. Kudva et al. [10] studied transfer learning using pre-trained AlexNet with labeled images. Most of the previous approaches extracted visual features using labeled data and subsequently the features were applied for classifying cancer. These methods heavily depend on the amount of annotated data.

While expanding on previous work for classifying cancer, this paper presents a new patch self-supervised learning method by extracting global context from unlabeled data. We also propose a weight-sharing model with an attention learning to focus on local features of the image with a small set of labeled data.

3. Data

We used a cervix image dataset maintained by the US National Cancer Institute (NCI) from 10,000 women [11]. The patients had multiple visits, and each visit consisted of multiple screening tests. The total number of images is 45,009. However, the number of labeled images within 1 year from the date of screening is only 978. Images have a resolution of approximately 2800×1900 pixels, and individual sizes vary. We used the unlabeled images for self-supervised learning to extract the global features, and the labelled images were used to classify cancer. The CIN grades (CIN0, CIN1, CIN2, CIN3, CIN4) are the ground truth and used to label the cervix images. CIN0 and CIN1 grades are labeled as normal, while CIN2+ are identified as abnormal cases. We divided the classes into positive (CIN2+) and negative (CIN0/CIN1). We split the images into two groups to classify cancer: one with labels for training, 80%, and the other for testing the classifier, 20%. The dataset classified by the presence of a label is shown in Table 1.

Table 1. The number of images with and without labels

Type	Number of images
Without label	44,031
With label	978
Total	45,009

4. Method

A high image resolution and a high expansion rate can minimize diagnostic errors. However, it is expensive to use high-resolution images due to the limitations of memory capacity and computational resources. Thus, we devised effective methods to utilize low-resolution images and introduced self-supervised learning to extract global features using unlabeled images. The pre-trained visual features from the self-supervised learning are used to classify cervical cancer. Finally, we present the cancer classifier that pays attention to the critical parts of the image.

4.1 Patch Self-supervised Learning

This work presents an effective method to extract global features using unlabeled images of low resolution. The crucial part of the image to diagnose cervix cancer is mostly located around the center. The irrelevant parts (e.g., instrument components from a speculum) are located near the periphery. As the model should learn such information, we applied the patch self-supervised learning method.

Since using a high-resolution image is expensive when training a deep learning model, we empirically resized the images to 512×512 pixels from the original sizes of approximately 2800×1900. Then, we divided each 512×512 image into nine patches (3×3) of 153×153 pixels with some space to avoid trivial solution and randomly shuffled the patches. Even though the possible combination of the nine patches is 9!, we randomly chose 1,000 cases from the combinations since similar combinations exist, such as [1,2,3,4,5,6,7,8,9] and [1,2,3,4,5,6,7,9,8], and the computational resources are limited. The goal is to learn the patch arrangement with 1,000 cases, like solving a 3×3 puzzle. The patches are passed through a self-supervised learning model based on ResNet50 that is shown in Fig. 1. The model consists of residual blocks, convolution blocks, and a pooling layer. There is a skip connection between the residual blocks.

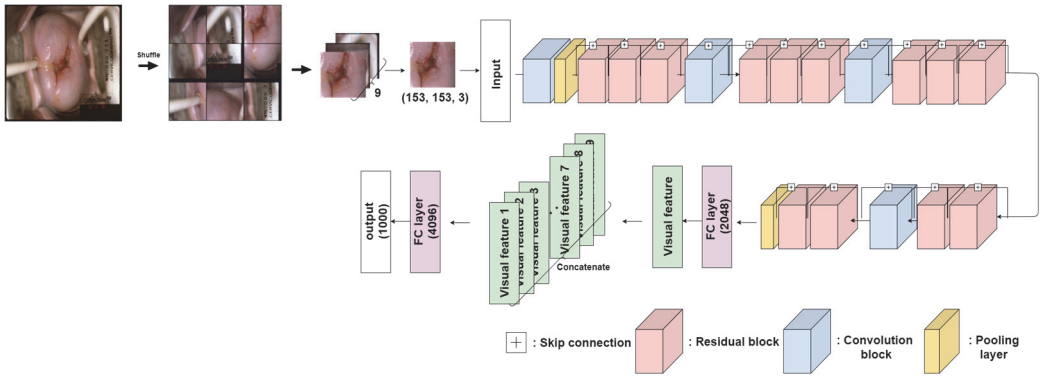


Fig. 1. The self-supervised learning model architecture based on ResNet50.

There is also a skip connection with a 1×1 convolution between the convolution blocks and residual blocks that increases the channel dimension. One of the nine patches is passed through the convolution layer, and the size is reduced through a max-pooling of the first pooling layer. Then, the feature is extracted through the blocks, and the last pooling layer uses an average pooling. Each patch's extracted features that passed the fully connected layer from the model is concatenated with the other patch's extracted features. From this patch arrangement learning, the model learns the global features.

4.2 Proposed Model

We train the cancer classifier model shown in Fig. 2 using the learned weights from the self-supervised learning to classify cervical cancer. The proposed model consists of residual blocks, convolution blocks, and a pooling layer based on ResNet50. Each residual block has two 2D convolution layers with ReLU and batch normalization, and a skip connection between the residual blocks is added. Each convolution block has two 2D convolution layers with ReLU and batch normalization. A skip connection is added for increased channel dimension between a convolution block and a residual block. The parts before the red boxes in Fig. 2 are the transferred knowledge learned from self-supervised learning. We divide the image into nine patches and classify cancer using the patches. Each patch is passed through the proposed model in Fig. 2. The convolution's weights are shared between the patches. Finally, each patch's feature, extracted through a fully connected layer (green rectangles in Fig. 2), is concatenated with other patch's features for classifying cancer.

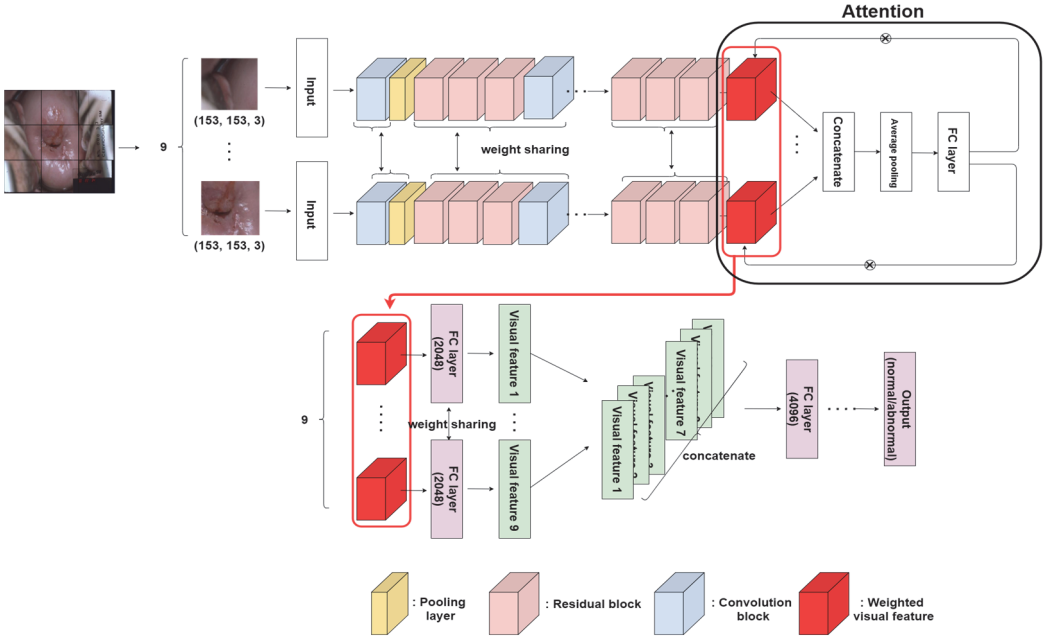


Fig. 2. The proposed model with attention based on ResNet50.

4.3 Attention Learning

To classify cancer, we need to pay attention to the critical parts of an image as is done with a human expert's diagnosis. The critical parts are usually near the center of an image, but for some images the

important parts could be off-center. Therefore, all the patches are passed through the model, sharing the model's weights, and then the extracted features are concatenated.

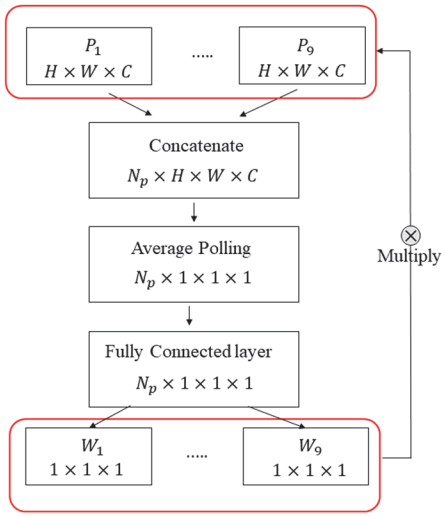


Fig. 3. Attention learning in detail. The output dimensions are given in each block.

To identify what patch should be paid attention to, we applied the attention learning method using an average pooling and a fully connected layer with sigmoid function, as shown in more detail in Fig. 3. The input patch size for attention learning is $H \times W \times C$ (H: height, W: width, C: channel). We stack all patches of size $N_p \times H \times W \times C$, where N_p is the number of patches. We apply a global averaging pooling to create the initial path weights of size N_p. From the fully connected layer with sigmoid function, we can learn what patch is important to classify cancer, and the weight of the patch is transformed to the size $N_p \times 1 \times 1 \times 1$. Thus, W_i , the weight of each patch is multiplied to reweight the original patch. In the learning, the output of the fully connected layer is multiplied by the extracted features, and the weighted visual features are derived, which are a red block in Fig. 2. The final fully connected layer reduces the dimension and extracts the final visual features for classifying cancer.

5. Results

The evaluation metrics are Accuracy, Specificity, Sensitivity, and Precision, as shown in Eqs. (1)–(4). In the Eqs. (1)–(4), TP and TN mean true positive and true negative, respectively. Also, FP and FN mean false positive and false negative, respectively.

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{TP + TN}}{\boldsymbol{TP + TN + FP + FN}} \quad (1)$$

$$\boldsymbol{Specificity} = \frac{\boldsymbol{TN}}{\boldsymbol{TN + FP}} \quad (2)$$

$$\boldsymbol{Sensitivity} = \frac{\boldsymbol{TP}}{\boldsymbol{TP + FN}} \quad (3)$$

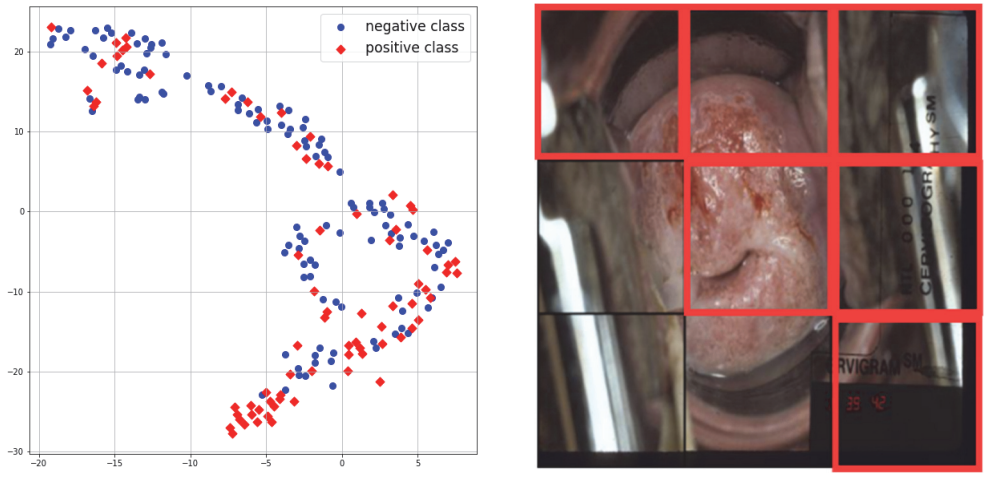
$$\boldsymbol{Precision} = \frac{\boldsymbol{TP}}{\boldsymbol{TP + FP}} \quad (4)$$

The experiments were carried out on the dataset with labels split into 80% training and 20% testing. We compared with three baseline results of state-of-the-art methods. The previous approaches have used specific manual annotations to use the RoI of an image, and those labels are not publicly available. Since the crucial region for diagnosing cancer is around the center of the image, to compare them with our study, we cropped the center of each image to 1000×1000 pixels and then applied the previous state-of-the-art works' methods. Table 2 shows a comparison of our results with previous works and baseline models. Overall, the results with self-supervised learning and attention learning show around 6% better performance than those without it. Compared with state-of-the-art results, our combined method shows 0.01% and 5.5% better performance in accuracy and sensitivity, respectively. In particular, sensitivity is an important measure in medical diagnosis. In Fig. 4(a), we show the feature visualization by the 2D T-SNE algorithm. The points of each class are clustered, but some points overlap. We show the impact of attention learning by visualizing each patch with higher weight than average in an image with a red border in Fig. 4(b) and an important patch is paid attention to. While the previous approaches require manual annotation to extract the RoI of the image, our method does not require such annotation and can use unlabeled data. Overall, our approach, which combines self-supervised learning and attention learning, achieves relatively good performance without needing a lot of annotated data.

Table 2. Experiment results comparison with state-of-the-art works (unit: %)

Experiment	Manual annotation	Use of unlabeled data	Accuracy	Specificity	Sensitivity	Precision
Xu et al. [2], 2017	✓	×	73.84	85.41	62.62	81.57
Vasudha & Juneja [4], 2018	✓	×	65.64	75.52	55.05	64.47
Alyafeai & Ghouti [3], 2020	✓	×	72.82	76.19	66.66	60.53
ResNet50	×	×	67.18	74.34	57.32	61.84
VGG16	×	×	62.56	65.33	53.33	31.58
ResNet50 + SSL + AL	×	✓	73.85	76.95	68.12	61.84
VGG16 + SSL+ AL	×	✓	68.72	76.36	58.82	65.79

SSL=self-supervised learning, AL=attention learning.



(a) (b)

Fig. 4. (a) Feature visualization and (b) attentive patch (red) in the image.

6. Conclusion & Future Work

This study proposes an approach in which self-supervised and supervised learning with attention using both unlabeled data and labeled data are combined. We improved performance over existing approaches by applying this method to a cervix image dataset. As many image analysis problems have limited annotated data, we expect the approach can be useful for similar problems. As future work, we plan to study an extension of the attention method to increase the accuracy and use external domain knowledge.

Acknowledgement

This research was supported by the MSIT (Ministry of Science, ICT), Korea (No. 2019-0-01599, High-Potential Individuals Global Training Program) supervised by the Institute for Information and Communications Technology Planning and Evaluation. We would like to express our gratitude to Dr. Mark Schiffman, Division of Cancer Epidemiology & Genetics, US National Cancer Institute, for allowing us to use one of the NCI datasets.

References

[1] M. Arbyn, E. Weiderpass, L. Bruni, S. de Sanjose, M. Saraiya, J. Ferlay, and F. Bray, "Estimates of incidence and mortality of cervical cancer in 2018: a worldwide analysis," *The Lancet Global Health*, vol. 8, no. 2, pp. e191-e203, 2020.

[2] T. Xu, H. Zhang, C. Xin, E. Kim, L. R. Long, Z. Xue, S. Antani, and X. Huang, "Multi-feature based benchmark for cervical dysplasia classification evaluation," *Pattern Recognition*, vol. 63, pp. 468-475, 2017.

[3] Z. Alyafeai and L. Ghouti, "A fully-automated deep learning pipeline for cervical cancer classification," *Expert Systems with Applications*, vol. 141, article no. 112951, 2020. https://doi.org/10.1016/j.eswa.2019.112951

[4] A. M. Vasudha and M. Juneja, "Cervix cancer classification using colposcopy images by deep learning method," *International Journal of Engineering Technology Science and Research*, vol. 5, pp. 426-432, 2018.

[5] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," in *Proceedings of the 6th International Conference on Learning Representations (ICLR)*, Vancouver, Canada, 2018.

[6] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proceedings of the 37th International Conference on Machine Learning (ICML),* 2020, pp. 1597-1607.

[7] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, WA, 2020, pp. 9729-9738.

[8] R. Gorantla, R. K. Singh, R. Pandey, and M. Jain, "Cervical cancer diagnosis using cervixnet: a deep learning approach," in Proceedings of *2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE)*, Athens, Greece, 2019, pp. 397-404.

[9] S. Mustafa and M. Dauda, "Evaluating convolution neural network optimization algorithms for classification of cervical cancer macro images," in *Proceedings of 2019 15th International Conference on Electronics, Computer and Computation (ICECCO)*, Abuja, Nigeria, 2019, pp. 1-5.

[10] V. Kudva, K. Prasad, and S. Guruvare, "Transfer learning for classification of uterine cervix images for cervical cancer screening," in *Advances in Communication, Signal Processing, VLSI, and Embedded Systems*. Singapore: Springer, 2020, pp. 299-312.

[11] ClinicalTrials.gov, "An innovative treatment for cervical precancer (UH3)," 2017 [Online]. Available: https://www.clinicaltrials.gov/ct2/show/NCT03084081.

Jinyeong Chae **https://orcid.org/0000-0002-4769-9889**

She received a B.S. degree in statistics and convergence software from Dongguk University in 2020. Since 2020, she is currently a M.S. student at the Department of Artificial Intelligence, Dongguk University, Korea. Her current research interests include computer vision, and knowledge-based reasoning.

Roger Zimmermann **https://orcid.org/0000-0002-7410-2590**

He received his B.S. degree in Informatik from the University of Applied Sciences and Arts, Northwestern Switzerland in 1986 and his M.S. and Ph.D. degrees in Computer Science from the University of Southern California, Los Angeles, California, USA, in 1994 and 1998, respectively. He is now a professor at the School of Computing at the National University of Singapore. His research interests include multimedia, networks and spatio-temporal data management.

Dongho Kim **https://orcid.org/0000-0003-3349-103X**

He received his B.S. degree in Computer Engineering from Seoul National University, Korea, in 1990 and his M.S. and Ph.D. degrees in Computer Science from the University of Southern California, Los Angeles, California, USA, in 1992 and 2002, respectively. He is now a professor at the Dongguk Institute of Convergence Education, Dongguk University, Korea. His research interests include artificial intelligence, distributed systems, networks, and security.

Jihie Kim **https://orcid.org/0000-0003-2358-4021**

She received a B.S. degree in computer science and statistics from Seoul National University in 1988, an M.S. degree in computer science and statistics from Seoul National University in 1990, and a Ph.D. degree in computer science from the University of Southern California in 1996. She is currently a professor at the Department of Artificial Intelligence in Dongguk University, Seoul, Korea. Her research interests include machine learning, NLP and knowledge-based reasoning.