

Vision Transformer-based Model for Gastric Cancer Detection and Classification using Weakly Annotated Histopathological Images

Tagne Poupi Theodore Armand*, Subrata Bhattacharjee**, Hyun-Joong Kim*, Ali Hussain*, Sikandar Ali*, Heung-Kook Choi**, and Hee-Cheol Kim*

*Institute of Digital Anti-Aging Healthcare, Inje University, Gimhae 50834, Republic of Korea

**Department of Computer Engineering, u-AHRC, Inje University, Gimhae 50834, Republic of Korea

poupiarmand2@gmail.com, subrata_bhattacharjee@outlook.com, play97509@gmail.com, alihussain.dream@gmail.com, sikandarshigri77@gmail.com, cschk@inje.ac.kr, heeki@inje.ac.kr

Abstract— Gastric Cancer (GC) is the fifth most diagnosed cancer worldwide. An early diagnosis is a hope for patients suffering from GC. A biopsy is a procedure that helps detect abnormal and suspicious areas to determine whether cancer cells are in the stomach. Tissue samples collected through biopsy are stained using Hematoxylin and Eosin (H&E) and digitalized through scanning to produce a whole slide image (WSI) needed for further analysis. Recently, most prognostics have proven effective using artificial intelligence techniques combined with related computer aid detection systems. This research used a vision transformer to detect and classify gastric cancer from weakly annotated tissue images. After acquiring normal and cancer histopathological samples, we applied the vision transformer (ViT) model for binary classification. We generalized our approach by performing region-based prediction on unannotated tissue samples. The proposed approach will ease diagnosis and support pathologists in decision-making.

Keywords— Gastric Cancer, Vision Transformers, Whole Slide Images, Weakly Annotated Images.

I. INTRODUCTION

Computer Aids Detection (CAD) uses various algorithms with different technologies to perform specific tasks such as pattern analysis, anomaly detection, or matches between data. CAD has proven effective and efficient for multiple applications in many industries. In healthcare, CAD using artificial intelligence has demonstrated outstanding capabilities [1]. Machine learning and artificial intelligence algorithms have been applied to detect and classify diseases using customized and pre-trained models [2]. The data collection process is carried out by medical personnel, after which data undergoes necessary preprocessing and must be fed to the model for training using specific algorithms according to the task to be accomplished. For instance, machine learning algorithms have demonstrated some capability in handling disease prediction [3-5].

Furthermore, deep learning methods have expanded the possibilities by offering chances to develop algorithms capable of improving diagnosis at optimal speed [6, 7]. One of the latest deep learning-based algorithms incorporating these features is Vision Transformers (ViTs). ViTs are an innovative approach

used to handle image analysis tasks using the transformer architecture. Though the earlier transformer proposed by Vaswani et al. [8] introduced the transformer architecture applicable to natural language processing tasks, vision transformers by Dosovitskiy et al. [9] extended their application to image processing tasks.

In medical imaging, various images (X-ray, CT, MRI, Ultrasound, WSI...) are used for disease diagnoses, prognoses, and treatments. Histopathology Whole Slide Images (WSI) are commonly used for analysis and are often obtained through biopsy. An organ is collected from the patient through biopsy and then stained using Hematoxylin and Eosin (H&E) to prepare the tissue for digitalization through a digital scanner. WSI is being processed using many algorithms, such as machine learning and deep learning algorithms, including CNN and vision transformers. Vision transformers have proven effective for many computer vision tasks, including medical image classification, object detection, and image segmentation. Ikromjanov et al. proposed a vision transformer-based model to detect prostate cancer [10]. The proposed model was built using histopathological images and could detect and grade prostate cancer according to the Gleason grading system using pre-trained ViTs and prostate WSI.

Similarly, Kollias et al. [11] proposed a Covid-19 diagnosis tool developed with ViTs. This model was able to detect region-based lesions on CT scan images. Zeid et al. [12] developed a vision transformer and CNN-based model known as compact convolutional transformer to achieve the classification of colorectal cancer using histopathology images. ViTs have also been influential in the segmentation tasks. An efficient Coronary artery segmentation was done by Ning et al. using vision transformers [13]. Their model consisted of three main modules: one for feature extraction, another for modeling the features to high-level texture features, and a last module for context aggregation. The experimental result demonstrated good performance with a more than 75% dice coefficient. Table 1 below shows more applications of vision transformers in medical imaging tasks.

TABLE 1: APPLICATION OF VISION TRANSFORMERS IN MEDICAL IMAGING

SN	Description	Archi.	Ref.
1	Classification of retail disease using fundus images and ViTs	Base	[14]
2	Multi-scale Hybrid Vision Transformer for Learning Gastric Cancer Histology	Hybrid	[15]
3	Applying ViTs for Breast ultrasound image classification	Base	[16]
4	MIL transformer-based model for classification of histological Whole Slide images	Hybrid	[17]
5	Multiclass Colorectal Cancer Histology Images Classification Using Vision Transformers	Base	[12]

In this research, we use the base vision transformer with weakly annotated areas of WSI to detect and classify gastric cancer. The main contribution of this work is to use weakly annotated regions of whole slide images and ViTs to efficiently classify gastric cancer histopathological images in order to support physician diagnostics and decision support systems.

II. DATASET

The histopathological samples used in this experiment were sourced from the GasHisSDB dataset [26], accessible at <https://gitee.com/neuhwm/GasHisSDB>. This dataset comprises images with diverse imaging characteristics and is publicly available for research. The histopathological samples were curated by four pathologists from Longhua Hospital, Shanghai University of Traditional Chinese Medicine, resulting in 560 cancerous images and 140 normal images, each with a resolution of 2048×2048 pixels. The dataset was prepared through H&E staining at a magnification of 20X using Nikon (Japan) and Olympus (Japan) Microscopes. These images were extracted from original microscopic biopsy images related to gastric lesions, specifically highlighting cancerous regions. The dataset has been well-annotated by medical experts. The example samples of the dataset are shown in Figure 1.

III. METHODS

In this research, we used 2048x2048 weakly annotated region extracted from a whole slide image of gastric cancer. The weak annotation provides region-level labels of patients with cancer or normal, used for disease detection and classification using the vision transformer model. To achieve our goal, we had to prepare the data and feed it into the vision transformer to train the model that will serve as a region-level classifier of the WSI patches, which will help to achieve the complete classification of any given gastric cancer WSI into the cancer and normal classes.

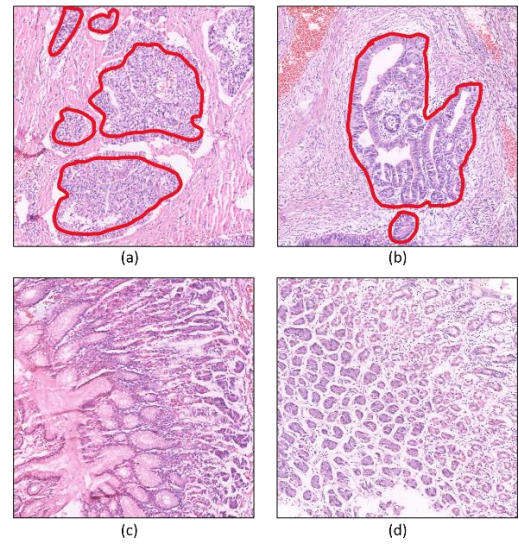


Figure 1. Sample images of the dataset. (a, b) Cancer samples with annotated regions. (c, d) Normal samples.

A. Data Preprocessing

For each region of interest (ROI) annotated by the pathologist, we conducted manual weak annotation for weakly supervised learning. Tissue samples of size 2048×2048 were generated from these weakly annotated ROIs. The patching process enhances scalability and facilitates efficient analysis by allowing precise examination of specific areas. Utilizing these patch images reduces computational requirements, resulting in more optimal learning, inferencing, and likelihood estimation times than directly manipulating whole slide images (WSIs). The preprocessing involves a series of operations on the patch images to prepare them for analysis. This includes removing the background, as illustrated in Figure 2, to gain insights into the image information.

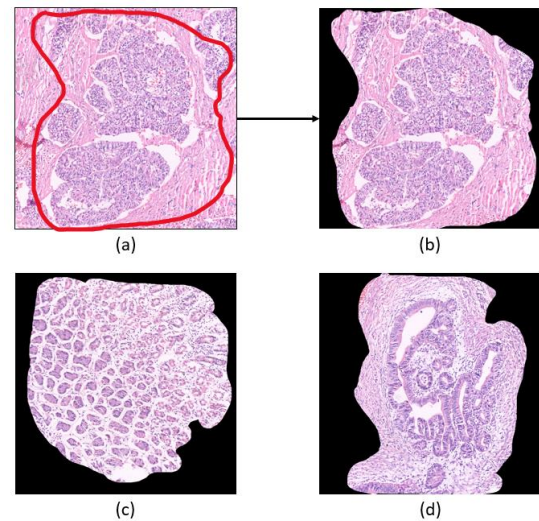


Figure 2. Preprocessed tissue samples. (a) Weakly annotated tissue sample. (b) Extracted region of interest from (a) by removing the background. (c) Example sample of normal tissue. (d) Example samples of cancer tissue.

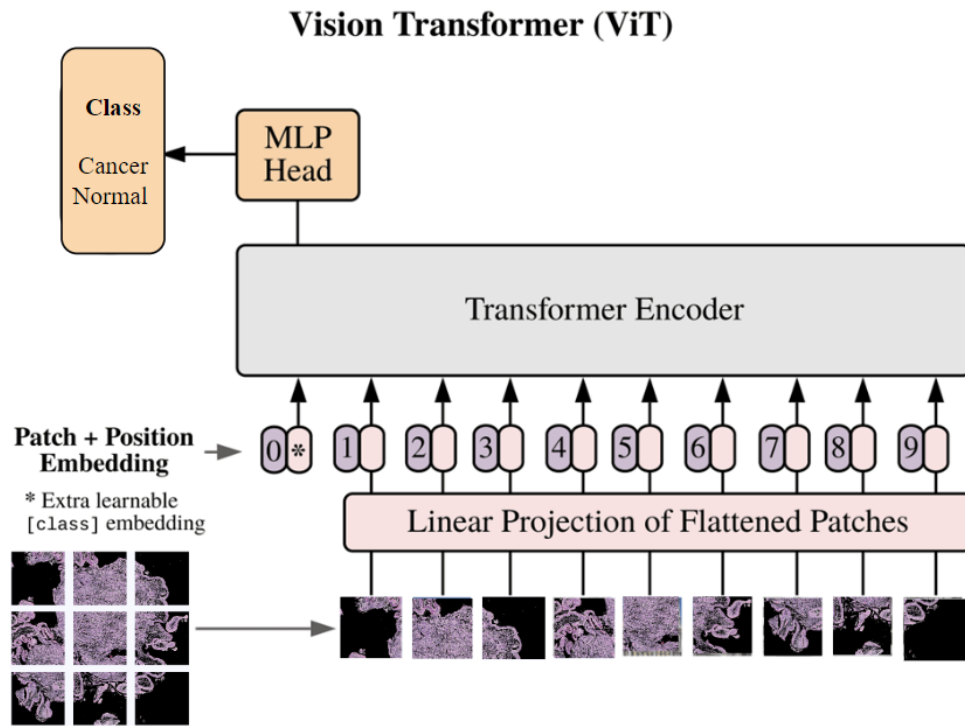


Figure 3. Architecture of Vision transformer.

B. Vision transformer

Vision transformers are deep learning models designed for computer vision-related tasks (detection, classification, and segmentation). The success of the transformers model in natural language processing inspired the vision transformer presented by Dosovitskiy et al. [9]. Traditionally, computer vision tasks were mostly achieved using CNN; nowadays, the vision transformers are a new approach recently gaining much attention from industry and academia. ViTs use a self-attention mechanism that allows the computer to understand the relationship between various parts of an image. It assesses the importance of the patch score and focuses on the most relevant information helpful for the model. The patch size defined in the original paper was set to 16×16 and can be changed. In this study, we considered a 16×16 patch size, generating 16,384 patches for a single input image of size 2048×2048 . The input images are tokenized and flattened; in this process, the 2-D patch of 16×16 is converted into a 1-D array of 256 tokens that will be input into the linear projection layer. The linear projection layer will transform the 1-D vector into lower-dimensional vectors while preserving relevant information. For any 1-D vector x_1 , the weight W and a bias b obtained during training produce an output expressed as $Wx_1 + b$ representing the lower dimensionality transform vector. The lower dimensionality vectors are the reduction in the number of features used to describe an object. This process benefits our algorithm by reducing noise, computational power, and memory requirements, thus increasing speed. Position embedding takes the lower dimensional vector as input and provides the position information to the transformer encoder. Unlike the original transformer, the ViTs don't have a decoder

part. The encoder transformer's first layer is the self-attention layer (Multi-Head Attention Layer), which captures dependencies between patches and enables the model to consider the global context. The output of each patch passes through a feed-forward network that captures the complex non-linear relationship between the patches. The next stage, at the final layer, is the classification layer, which matches the output of the transformer layer (Multilayer Perceptron head) and the desired output (cancer/normal classification in our case).

IV. RESULTS AND DISCUSSION

In this paper, we focused on the various hospital cohorts for training and testing using the deep learning framework, the ViT model, to classify region-level whole slide images of cancer and normal. The ViT model used in the study provided some sound results in classifying region-level histopathologic images containing aggressive and nonaggressive tumor cells. The experiment was also carried out using CNN-based algorithms, notably ResNet50, DenseNet121, and EfficientNetB1, which were chosen for their effectiveness in medical image classification tasks.

The ViT-16 model prompted a test accuracy of 85.9%, outperforming all other models. The overall performance is of all the models are shown in Table 2. The performance metrics used for model evaluation are accuracy, precision, recall, and f1-score, computed as shown in equations 1-4. Moreover, we split the dataset into training and validation with an 80:20 ratio while reserving 40 independent weakly annotated slides for testing purposes.

TABLE 2: OVERALL CLASSIFICATION PERFORMANCE OF THE ViT MODEL

Class	Accuracy	Precision	Recall	F1-score
ResNet50	70.59	0.68	0.63	0.64
DenseNet121	74.51	0.75	0.73	0.74
EfficientNetB1	78.71	0.76	0.81	0.76
ViT	85.9	0.77	0.85	0.78

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

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

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

$$F1 - score = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (4)$$

The results of our ViT model show that it can achieve a competitive performance and be effectively used in classifying cancer and normal gastric cancer cases. From the pathological perspective, cancer cells are graded based on their level of differentiation. Well-differentiated cells closely resemble normal, healthy cells, while poorly differentiated cells look very different from normal cells and tend to be more aggressive. Poorly differentiated adenocarcinoma is associated with a higher grade and is often more aggressive in its growth and spread. The proposed model aligns with the pathological analysis, and some observations can be made on the output image. Our model was trained with the most suitable hyperparameters with a few epochs set to $n = 20$. We evaluated the performance of our model with a variety of metrics. Figure 4 illustrates the model learning graphs of accuracy and loss that give insights into the training process and performance. The model's efficiency can also be observed from the confusion matrix shown in Figure 5. From the confusion matrix, we can see that there are a few mispredictions, and it is because of the similarity samples of cancer and normal; as a result, the model did not perform well for some specific images.

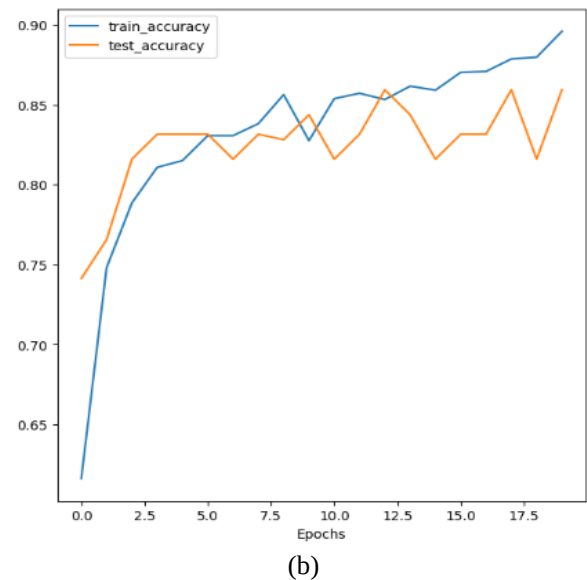
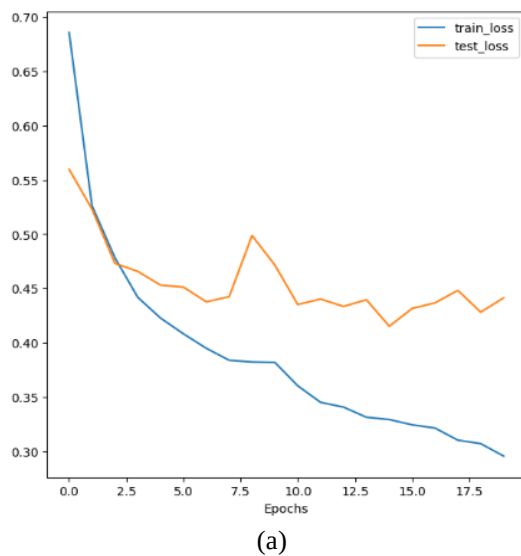


Figure 4. Learning curves for the ViT model. (a) Loss curve; (b) Accuracy curve.

		Predicted Label	
		Cancer	Normal
Actual Label	Cancer	10	1
	Normal	8	31

Figure 5. Confusion matrix of the ViT-based model



The model we used in the work showed astounding results for the cancer class by providing an overall accuracy and recall of 0.85 and 0.85, respectively. The ViT-based classification models are very much necessary for classifying region-level WSI images which can assist pathologists in the identification of tumor and non-tumor cells in H&E histopathological samples. There are many existing papers on weakly supervised approaches where they performed patch-level multiple-instance learning (MIL) for binary classification. However, we presented a state-of-the-art method for weakly supervised learning by feeding the global features to the model extracted from region-level weakly annotated gastric cancer samples. This approach was carried out to bypass the information loss that occurs in other patch-based MIL approaches.

V. CONCLUSIONS

In conclusion, this paper proposes a method to detect and classify gastric cancer using ViTs and weakly annotated histopathological images. Unlike the CNN approaches that use low-level to high-level feature extractor mechanisms for classification with effects on computational and statistical efficiency, the vision transformer uses the attention mechanism to overcome most CNN challenges while reducing dependencies powered by its parallel processing architecture. The obtained results showed the potential to assist the pathologist in decision-making. Region-level segmentation can be further used to compare our approach with fully supervised learning-based methods.

ACKNOWLEDGMENT

This research was supported by a grant from Korea Health Technology R&D Project through the Korea Health Industry Development Institute validation (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (Grant No: HI21C0977)

REFERENCES

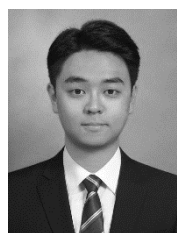
- [1] Shiraishi, Junji, et al. "Computer-aided diagnosis and artificial intelligence in clinical imaging." *Seminars in nuclear medicine*. Vol. 41. No. 6. WB Saunders, 2011.
- [2] Ahsan, Md Manjurul, Shahana Akter Luna, and Zahed Siddique. "Machine-learning-based disease diagnosis: A comprehensive review." *Healthcare*. Vol. 10. No. 3. MDPI, 2022.
- [3] Ansari, A.Q.; Gupta, N.K. Automated diagnosis of coronary heart disease using neuro-fuzzy integrated system. In Proceedings of the 2011 World Congress on Information and Communication Technologies, Mumbai, India, 11–14 December 2011; pp. 1379–1384.
- [4] Levey, A.S.; Coresh, J. Chronic kidney disease. *Lancet* 2012, 379, 165–180.
- [5] Vidushi, A.R.; Shrivastava, A.K. Diagnosis of Alzheimer disease using machine learning approaches. *Int. J. Adv. Sci. Technol.* 2019, 29, 7062–7073.
- [6] Serag, Ahmed, et al. "Translational AI and deep learning in diagnostic pathology." *Frontiers in medicine* 6 (2019): 185.
- [7] Litjens, Geert, et al. "Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis." *Scientific reports* 6.1 (2016): 26286.
- [8] Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).
- [9] Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." *arXiv preprint arXiv:2010.11929* (2020).
- [10] Ikromjanov, Kobiljon, et al. "Whole slide image analysis and detection of prostate cancer using vision transformers." *2022 international conference on artificial intelligence in information and communication (ICAIC)*. IEEE, 2022.
- [11] Kollias, Dimitrios, et al. "Mia-cov19d: Covid-19 detection through 3-d chest ct image analysis." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021.
- [12] Zeid, Magdy Abd-Elghany, Khaled El-Bahnasy, and S. E. Abo-Youssef. "Multiclass colorectal cancer histology images classification using vision transformers." *2021 tenth international conference on intelligent computing and information systems (ICICIS)*. IEEE, 2021.
- [13] Ning, Yang, et al. "Cac-emvt: Efficient coronary artery calcium segmentation with multi-scale vision transformers." *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2021.
- [14] Yu, Shuang, et al. "Mil-vt: Multiple instance learning enhanced vision transformer for fundus image classification." *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VIII 24*. Springer International Publishing, 2021.
- [15] Oh, Yujin, et al. "Multi-Scale Hybrid Vision Transformer for Learning Gastric Histology: AI-Based Decision Support System for Gastric Cancer Treatment." *IEEE Journal of Biomedical and Health Informatics* (2023). Gheflati, Behnaz, and Hassan Rivaz. "Vision transformers for classification of breast ultrasound images." *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2022.
- [16] Shao, Zhuchen, et al. "Transmil: Transformer based correlated multiple instance learning for whole slide image classification." *Advances in neural information processing systems* 34 (2021): 2136–2147.



Tagne Poupi Theodore Armand was born in Cameroon and received an MSc in Information Systems and Networking at ICT University USA, Cameroon Campus, in 2021. He is a Ph.D. research scholar at the Institute of Digital Anti-aging and healthcare at Inje University. His research interest field includes image processing, focusing on medical image analysis, deep learning, machine learning, metaverse, and digital technology.



Subrata Bhattacharjee received his B.S. in information technology (IT) from the University of Derby, UK 2016. He is pursuing his M.S. leading Ph.D. from the Graduate School of Computer Engineering, Inje University, Korea. He is a Research Scholar and Teaching Assistant at the Medical Image Technology Laboratory (MITL) at Inje University. His research interests include multimodal medical imaging, tissue and cell image analysis, digital pathology, image reconstruction, cell and gland segmentation, machine learning, and deep learning.



Hyun Joong Kim was born in the Republic of Korea in 1994. He received his Bachelor's Degree in Nursing from Yonsei University in 2019. He is currently pursuing a Master's Degree majoring in artificial intelligence in healthcare at the Institute of Digital Anti-aging and healthcare, Inje University. His research activity is concentrated on Image Processing, Deep Learning, and Machine Learning.



Ali Hussain received his BSCS degree in computer science from Government College University Faisalabad (GCUF), Pakistan, in 2019. Furthermore, he got his master's from InjeUniversity South Korea in the Department of Digital Anti-Aging Healthcare. Currently doing PhD from the same department. His research interests include artificial intelligence, data science, Big data, machine learning, deep learning, Computer vision, reinforcement learning, and medical imaging.



Sikandar Ali received his B.E. degree in Computer Engineering from Mehran University of Engineering & Technology, Pakistan. He got his MS from the Department of Computer Science from Chungbuk National University, the Republic of Korea. Furthermore, he is now a Ph.D. candidate at Inje University South Korea in the Department of Digital Anti-Aging Healthcare. His research interests include artificial intelligence, data science, big data, machine learning, deep learning, reinforcement learning, Computer vision, and medical imaging.



Heung Kook Choi received his Bachelor of Engineering and Master of Engineering from Linköping University, Sweden, in 1988 and 1990, respectively. In 1996, he pursued his Ph.D. at the Centre for Image Analysis in Computerized Image Analysis from Uppsala University, Sweden. He has been a Professor at Inje University from 1997 to 2020. He is an Emeritus Professor of Computer Engineering at Inje University and operates the Medical Image

Technology Laboratory (MITL). He is the author of eight books, with more than 400 articles and more than 40 inventions. His research interests include computer graphics, multimedia, image processing, and analysis.



Hee-Cheol Kim received his BSc at the Department of Mathematics, MSc at the Department of Computer Science at SoGang University in Korea, and Ph.D. in Numerical Analysis and Computing Science at Stockholm Stockholm University in Sweden in 2001. He is a professor at the Department of Computer Engineering and Head of the Institute of Digital Anti-aging Healthcare Inje University in Korea. His research interests include machine learning, deep learning, Computer vision, and medical informatics.