IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 29 May 2024, accepted 5 July 2024, date of publication 15 July 2024, date of current version 24 July 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3428572

RESEARCH ARTICLE

Neural Network-Based Pipeline With High-Resolution Feature Enhancement and Low-Resolution Feature Preservation for Automated Treatment Decision of Graves' Orbitopathy Patients

SANGHYUCK LEE[1], MOHD ASYRAF ZULKIFLEY[2], (Member, IEEE), JEONG KYU LEE[3], AND JAESUNG LEE[1]

[1]Department of Artificial Intelligence, Chung-Ang University, Seoul 06974, Republic of Korea
[2]Department of Electrical, Electronic and Systems Engineering, Universiti Kebangsaan, Bangi 43600, Malaysia
[3]Department of Ophthalmology, Chung-Ang University College of Medicine, Chung-Ang University Hospital, Seoul 06973, Republic of Korea

Corresponding authors: Jeong Kyu Lee (lk1246@cau.ac.kr) and Jaesung Lee (curseor@cau.ac.kr)

This work was supported in part by the Chung-Ang University Research Grants in 2023; in part by the National Research Foundation of Korea (NRF) through the Korean Government (MSIT) under Grant NRF-2021R1A2C1011351; and in part by the Institute of Information and communications Technology Planning and Evaluation (IITP) through the Korean Government (MSIT), Artificial Intelligence Graduate School Program (Chung-Ang University) under Grant 2021-0-01341.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Institutional Review Board of Chung-Ang University Hospital.

ABSTRACT Graves' orbitopathy is an inflammatory disorder that causes changes in different structures close to the eye. Accurate and consistent diagnoses are essential to improve the quality of life for Graves' orbitopathy patients. To this end, a number of studies on Graves' orbitopathy have been conducted based on neural networks recently. However, treatment decision methods based on neural networks have been much less addressed. This study aims to propose an effective deep neural network-based diagnosis method that makes treatment decisions for Graves' orbitopathy patients. Specifically, the proposed method adopts a high-resolution feature enhancement and low-resolution feature preservation strategy focusing on the following points. First, the loss of detailed spatial information during the alignment of pixel spacing in computed tomography images leads to a decrease in performance. Thus, we preserve the detailed information of the images through high-resolution resampling. Second, existing studies lack sophistication in network design. The baseline network was improved by four modifications. Finally, resizing and coarse cropping cause learning instability. Thus, we adopt padding and fine-grained cropping. Our empirical study shows that the proposed method outperforms two existing neural network-based Graves' orbitopathy diagnostic pipelines achieving an average area under the receiver operating characteristic curve of 0.793, accuracy of 0.699, F1 score of 0.416, sensitivity of 0.723, and specificity of 0.694 in five repetitive experiments. Furthermore, in-depth analysis provides several future research directions in computed tomography preprocessing and deep neural network design. The source code for the proposed model is available at https://github.com/tkdgur658/GOTDNet.

INDEX TERMS Graves' orbitopathy, computed tomography, deep neural network, treatment decision.

The associate editor coordinating the review of this manuscript and approving it for publication was Sangsoon Lim.

I. INTRODUCTION

Graves' orbitopathy (GO) is an orbital inflammatory disorder that occurs in association with autoimmune thyroid disease [1]. Patients with GO can experience reduced life quality due to several conditions, such as visual impairment, deformed appearance, and declined emotional health [2]. Moreover, GO primarily occurs in patients with hyperthyroidism but can also coexist with euthyroidism, subclinical hypothyroidism, or overt hypothyroidism [3]. For the systematization of treatment plans of GO patients, various taxonomies have been studied in the past years [4]. For example, the European Group on Graves' Orbitopathy categorized the severity of GO into three groups, and suggested different treatment plans for each group, taking into consideration the activity of GO [5]. Although numerous treatment methods have been explored, there is still a strong emphasis on the need for consistent treatment guidelines because the severity and activity of GO require different approaches [6].

At the same time, imaging techniques such as computed tomography (CT) play a crucial role in evaluating orbital alterations to understand the progression of GO and to assist in surgical planning [7]. Integrating pertinent input data along with meticulously crafted machine learning algorithms into diagnostic processes can assist physicians in making more accurate and consistent decisions [8]. Moreover, GO diagnosis based on orbital imaging is often time-consuming in practice because clinical decisions should be made after observing changes in various anatomical structures related to GO [9]. Thus, several studies of GO diagnosis have recently been conducted based on deep neural networks (DNNs) [10] for different objectives, such as the differentiation of GO patients from normal cases [11], severity evaluation [12], and activity evaluation [13], [14]. Nevertheless, the application of DNN-based methods for treatment decision making is still a relatively unexplored area [10], [15].

This study aims to develop an effective DNN-based diagnostic method for making treatment decisions for patients with GO. To this end, we design an effective diagnostic pipeline based on a high-resolution feature enhancement and low-resolution feature preservation strategy. In the preprocessing phase, we employ high-resolution resampling and fine-grained cropping strategies to preserve detailed spatial information from high-resolution images while alleviating less important region training. In network design, the deep stem strategy is utilized to extract high-level features in the high-resolution processing stage. Moreover, additional max-pooling, residual stage, and average-pooling enhance preservation of spatial information during feature extraction, thereby mitigating the rapid loss of important features. Finally, the bottleneck is designed to preserve local information by adding nonlinearity while reducing the receptive field. Specifically, we focus on the following points to construct a sophisticated diagnosis pipeline:

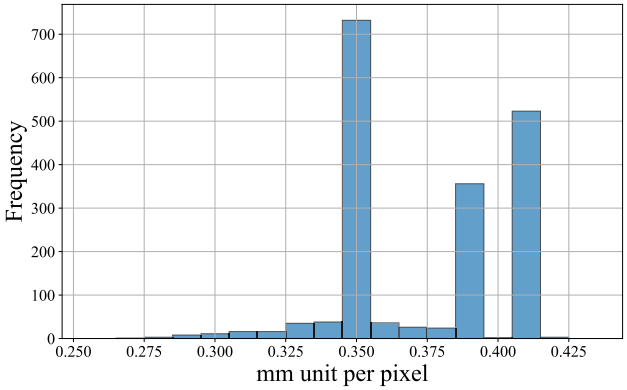


FIGURE 1. Distribution of physical units per pixel in the axial plane image from the CT dataset utilized in this study. A pixel spacing of $0.35mm^3$ constitutes the majority.

1) The loss of fine spatial information in CT images arising from resampling can lead to a decline in performance. For example, existing studies [14] normalize high-resolution CT images with a pixel spacing of approximately $0.34mm^3$ to a significantly lower resolution with a pixel spacing of $1.00mm^3$. Such approaches substantially diminish image resolution, thereby compromising intricate spatial information. However, the proposed methods alleviate this information loss by normalizing to a pixel spacing of $0.35mm^3$ to preserve the detailed spatial information. Fig. 1 delineates the distribution of physical units per pixel before resampling within the axial plane imagery of the CT dataset used in this study.
2) Although existing studies [11], [14] have shown potential in the use of DNNs on GO diagnosis, they lack a comprehensive theoretical or experimental basis for network design. In contrast, we analyze several network designs pertinent to improving predictive performance. To enhance the extraction of fine CT features, the proposed DNN employs a deep stem strategy, which offers a larger receptive field and greater non-linearity during the extraction of high-resolution features. Furthermore, to preserve extracted high-resolution features with reduced spatial distortion at lower resolutions, one $3 \times 3 \times 3$ convolution is replaced with two $1 \times 1 \times 1$ convolutions per convolution block, and stride convolutions are replaced with average pooling.
3) Resizing and coarse cropping of existing studies [11], [14] can cause learning instability. Spatial axis misalignment due to resizing can cause performance degradation, as a deformation of the appearance of structures is one of the symptoms of GO. Performing coarse cropping can include regions that are less relevant to GO during the decision of model. A network trained on coarse cropping images may suffer from regions that are less relevant to GO.

The proposed method mainly consists of two stages: data preprocessing and DNN architecture. The preprocessing of the proposed method is distinguished from existing studies by incorporating high-resolution resampling, fine-grained cropping, and no-resizing. In terms of DNN architecture, the proposed DNN was improved from an existing GO diagnostic DNN [11] based on a high-resolution feature enhancement and low-resolution feature preservation strategy. Specifically, we added max-pooling and a residual stage to the baseline network for effective high-resolution CT data processing. Furthermore, the bottleneck blocks and average-pooling were adopted for low-resolution feature preservation. Finally, deep stem strategy was exploited to extract better high-resolution features.

The main contributions of our study are as follows:

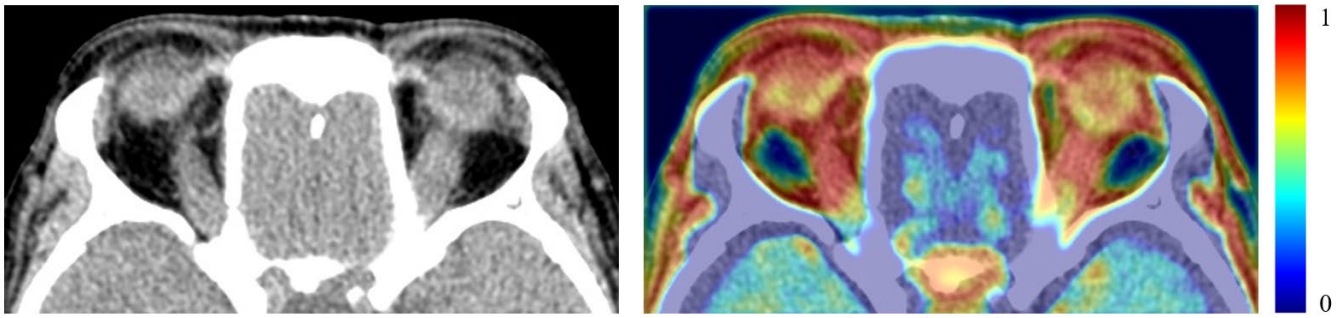
1) This study introduces a DNN-based diagnostic method for making treatment decisions for patients with GO, demonstrating superior performance over two existing methods across five evaluation metrics in experiments involving a cohort of 1,832 patients at Chung-Ang University Hospital.
2) This study tackles the problems of existing DNN-based GO studies in terms of preprocessing, such as the information loss from resampling and resizing, and learning confusion from GO-irrelevant regions. Thus, the proposed preprocessing includes high-resolution resampling and fine-grained cropping to alleviate information loss in CT data and focus on GO-relevant regions.
3) To design a DNN specialized for the proposed preprocessing, the proposed DNN exploits high-resolution feature enhancement and low-resolution feature preservation strategy including different architectural modifications.

The remainder of this paper is organized as follows: Section II describes the literature review. Section III presents the materials and methods. Section IV describes the experimental results. Section V presents an in-depth analysis of this study.

II. RELATED WORK

In recent years, a number of DNN-based diagnostic methods have been studied based on different approaches in terms of CT preprocessing, DNN designs, and training strategies to address different medical objectives [16]. For drowning diagnosis, a 2.5D method can be used for converting 3D CT data into 2D images to train 2D DNNs [17]. The CT images were processed by removing the background and then input into Inception-ResNet-V2. A computer-aided diagnosis framework based on multi-channel three-dimensional DNN was developed for lesion diagnosis [18]. To exploit the energy-enhanced tissue features, each energy image was used as one input channel for a multi-channel input convolutional neural network. Non-contrast CT images can be used for early diagnosis of pancreatic cancer based on a multiple-instance-learning framework designed to extract fine-grained pancreatic tumor features [19]. A patch-level feature extraction was performed to obtain local fine-grained features before inputting them into the graph neural network. Different medical images, which can have unique characteristics, including intensity variation, scale variation, and location of interest, can improve diagnostic performance when applied before feeding images into a DNN [20]. The diverse DNN-based diagnostic methods [16], [21], including the above examples, can indicate that GO studies should further employ a wide range of methods [10], [15].

For diagnosing GO patients, several studies [11], [12], [13], [14], [22], [23] have explored various DNN-based diagnostic pipelines. To assess GO activity, magnetic resonance imaging (MRI) scans from 108 patients were analyzed using two DNN models adapted from the VGG and ResNet architectures, respectively [13]. In the MRI scans, the orbital regions were extracted at a uniform size before being randomly cropped to diversify the training dataset. Although MRI provides enhanced contrast resolution over CT scans, its application can be limited due to higher costs and the limited availability of training datasets. Although CT data is relatively suitable in terms of cost for training data expansion, labor-intensive data preprocessing procedures may be difficult to apply in clinical settings [24]. To diagnose the severity of GO, multiple convolutional blocks were employed to analyze information across the axial, coronal, and sagittal planes [12]. Extracting latent variables from these three distinct views enabled the model to effectively learn 3D spatial information. However, the manual rectangular cropping procedure of all orbital regions for three views can be considered a labor-intensive process. Although a large number of segmented images can be a promising approach for improving predictive performance [23], [25], labor-intensive issues for segmentation data preparation should be addressed for clinical field application. The ResNet-VGG pipeline was designed to detect enlarged extraocular muscles, which are considered the main symptom of GO [22]. The use of segmentation neural networks can be a labor-intensive solution, but it also requires a large amount of mask data for learning. Thus, this study focuses on the 3D region-of-interest (RoI) cropping strategy, which can be used at a relatively lower cost than the segmentation strategy [26]. A modified 3D-ResNet was developed to differentiate GO patients from healthy subjects using 1,435 CT images [11]. These images were cropped to a rectangular shape to display both the orbital bone and the eyeball in the sagittal view. However, this cropping approach inadvertently captured many areas not related to GO diagnosis, negatively impacting the predictive performance of DNN. Finally, the activity diagnostic performance of a multi-channel convolutional neural network can be enhanced by concurrently inputting orbital and single-photon emission CT images [14]. However, resampling to pixels occupying 1.00 mm^3 may represent a loss of information in CT data, which is too low-resolution compared to the original CT data.



(a) Input CT image (b) Grad CAM image

FIGURE 2. Visualization of an inference result of existing DNN-based GO diagnostic pipeline [11] including preprocessing and DNN archtecture. The left figure demonstrates an image at a specific depth location of input CT image, and the right figure depicts the Grad CAM score overlay on the left figure. Red indicates regions that contributed significantly to model predictions, and blue indicates regions that did not. The figure demonstrates that regions outside the orbit, which are less related to GO, affect the model decision.

Our brief review shows that DNN-based GO studies adopt different data preprocessing pipelines and model architectures. Although segmentation-based procedures and cropping procedures have their pros and cons, cropping-based methods are promising for actual clinical applications. However, in existing studies, the learning of DNN can be interrupted by structures unrelated to GO [11] or information can be lost during the low-resolution resampling process [14]. In this study, we propose a new data preprocessing pipeline and DNN architecture to devise a DNN-based diagnostic pipeline for treatment decisions of GO patients.

III. MATERIALS AND METHODS

The Institutional Review Board (IRB) of Chung-Ang University Hospital approved this study (IRB No. 2303-015-19462), and the requirement for informed consent was waived considering its retrospective design. This study was conducted in accordance with the ethical standards outlined in the Declaration of Helsinki.

A. MOTIVATION

Although orbital CT images can be acquired with higher resolution [27], [28], [29], several existing DNN-based GO methods resample the original CT images using 1.00 mm^3 pixel spacing [11], [14]. Meanwhile, various DNN studies show that high-resolution images can lead to better predictive performance [30], which indicates that the performance of GO diagnosis can be improved through a high-resolution CT resampling process. Moreover, existing cropping strategies that cover large regions of CT images may adversely affect model learning from unrelated GO regions (Fig. 2.); thus, fine-grained cropping is required in preprocessing. In addition, resizing approaches that distort the aspect ratio of images can lead to performance degradation. In this paper, we propose improved preprocessing guidelines, including high-resolution resampling and RoI cropping to alleviate information loss in CT data and focus on important regions. Then, the baseline network [11] is equipped with several modifications based on a high-resolution feature enhancement and low-resolution feature preservation strategy for processing high-resolution CT images effectively. Specifically, four network modifications were considered for effective dimensional compression, including an additional max-pooling layer, a residual stage, bottleneck blocks, and a deep stem stage. The proposed method, including data preprocessing and network, is demonstrated in Fig. 3.

TABLE 1. Subject characteristics.

Characteristics	Positive	Negative	*p*-value
Number of subjects	305	1,527	
Age (years, mean ± standard deviation)	43.89 ± 12.98	35.72 ± 11.87	<0.001
Sex (male/female)	88/217	316/1,211	<0.005

B. PARTICIPANTS

We obtained orbital CT scans (Philips Brilliance 256 Slice CT, Philips Healthcare Systems, Andover, MA, USA) without contrast from patients diagnosed with GO from March 2004 to November 2022. A total 1,832 CT scans from an equal number of GO patients were included in this study. Each CT scan was classified as either positive, indicating that treatment was recommended, or negative, indicating that treatment was not recommended. GO patients were diagnosed according to Bartley and Gorman's criteria [33]. GO activity and severity were evaluated according to the standardized criteria recommended by the European Group on Graves' Orbitopathy [5]. Patients with active, moderate-to-severe GO were recommended to proceed with treatment, which included intravenous (IV) steroid treatment. Two ophthalmologists, each with more than five years of experience in oculoplasty, were blinded to the CT results and made decisions regarding whether to recommend treatment. The number of positive cases was 305 (88 males, 217 females), and the number of negative cases was 1,527 (316 males, 1,211 females) (p-value<0.005). The mean ages of positive and negative cases were 43.89 ± 12.98, and 35.72 ± 11.87,

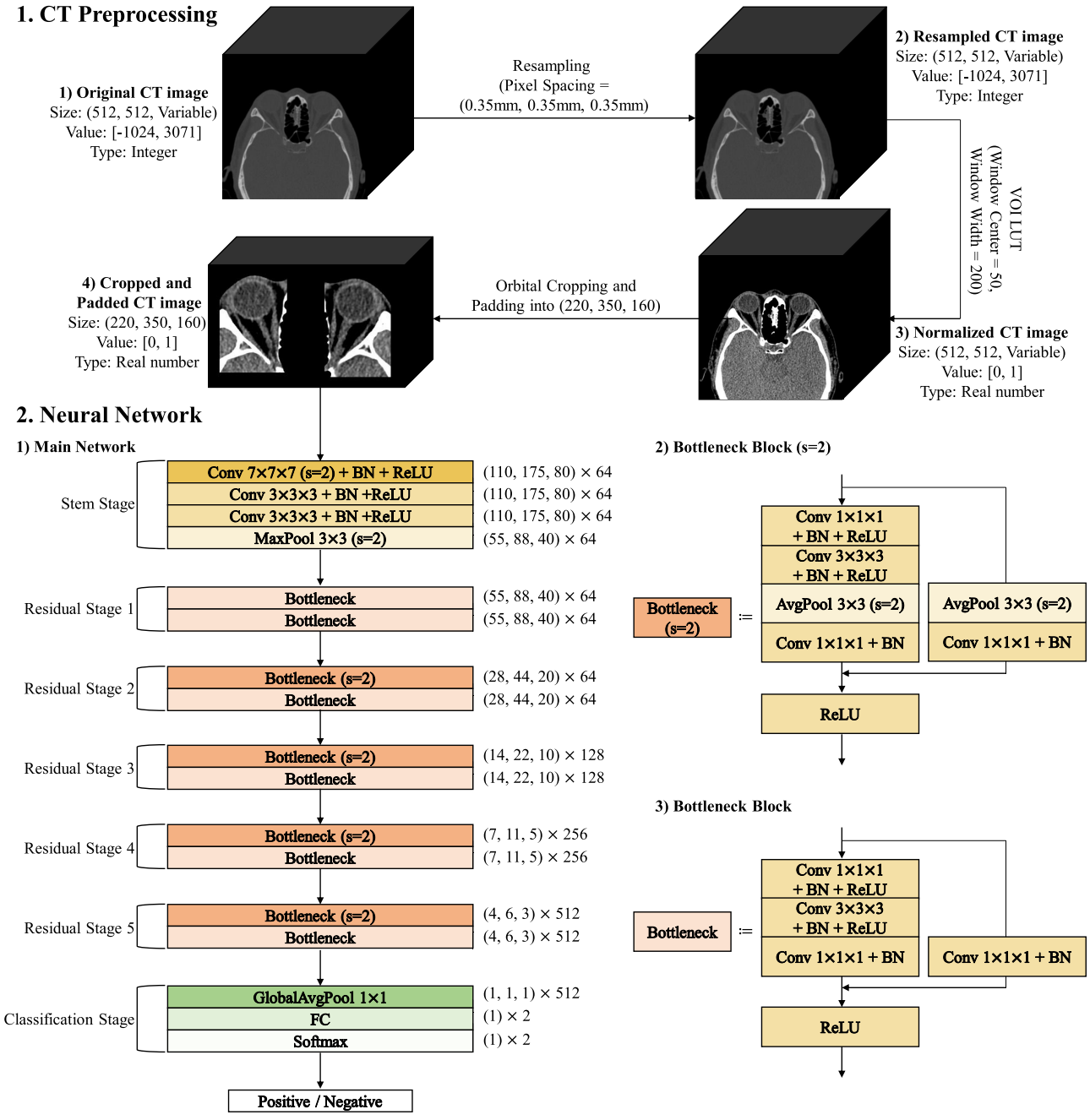


FIGURE 3. A schematic overview of the proposed method. First, the CT images were resampled based on B-spline interpolation with 0.35 mm^3 pixel spacing. Then, the images were normalized into values from zero to one by VOI-LUT function. The normalized CT image is cropped to the orbital region and padded with zero for having a fixed pixel size (220, 350, 160) for DNN. In the figure of the neural network, Conv stands for convolution layer. BN and ReLU denote batch normalization [31] and rectified linear unit [32], respectively.

respectively (p-value<0.001). Patients below the age of 18, with a previous history of orbital surgery, orbital tumor, blowout fracture, idiopathic orbital inflammation, those who received IV steroid treatment or radiation therapy at the time of the CT scan, or those with incomplete CT scans were excluded. The demographic information for the patients is described in Table 1.

C. DATA PREPARATION AND PROCESSING

Each CT image was resampled into 0.35 mm^3 with B-spline interpolation to obtain consistent pixel sizes across different CT scans. The 0.35 mm^3 pixel spacing represents a more than two-fold increase in resolution compared to the 1.00 mm^3 pixel spacing of the existing methods, resulting in significantly higher image quality. Then, each pixel value in

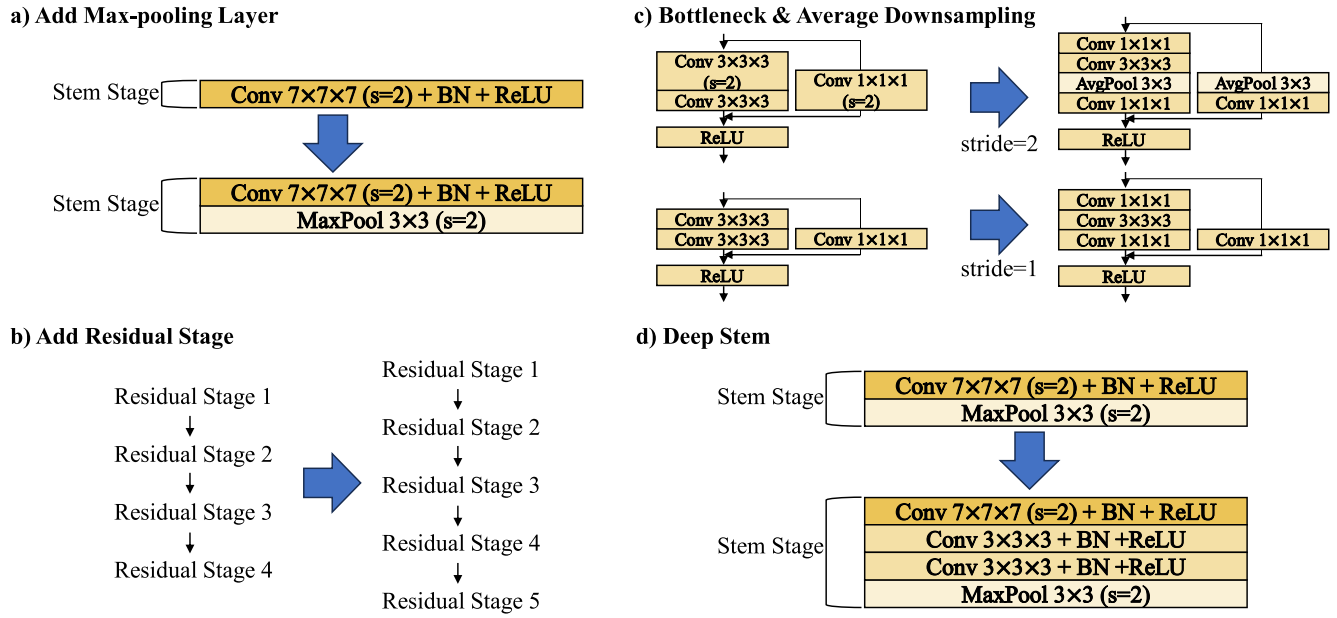


FIGURE 4. DNN architectural changes from the baseline [11] to the proposed model. a) First, a max-pooling layer is added to the stem stage of the baseline. b) Then, an additional residual stage is attached. c) Basic blocks of the baseline are replaced with bottleneck blocks. When s = 2, average-pooling is performed instead of strided convolution. d) Finally, two 3 × 3 × 3 convolutions are added to the stem stage.

CT images was normalized to a range of zero to one using the values of interest look-up table (VOI LUT) operation, with a window center of 50 and a window width of 200. To mitigate the confusion of training from GO-irrelevant regions, fine-grained orbital cropping was performed based on two rectangular boundaries containing both orbits. Before being inputted into the networks, the CT images were zero-padded to a fixed size of (220, 350, 160), avoiding the resizing approach used in existing methods [11].

D. DEEP NEURAL NETWORK

The proposed network adopts the high-resolution feature enhancement and low-resolution feature preservation strategy for processing the high-resolution CT images effectively. Specifically, the proposed network exhibits architectural changes using four steps, from the baseline network [11] as demonstrated in Fig. 4. First, the incorporation of max-pooling layers enhances feature extraction for high-resolution images, mitigating the potential for drastic information loss during global average pooling before the final FC layer. Second, an additional residual stage enhances the feature extraction for low-resolution features. Then, Bottleneck and average down-sampling facilitate the preservation of low-resolution features. Bottleneck alleviates spatial information distortion while improving non-linearity by replacing a $3 \times 3 \times 3$ kernel with two $1 \times 1 \times 1$ kernels. In addition, average-pooling entails a lower risk of information loss compared to strided convolution. Finally, the deep stem strategy enhances the learning ability from high-resolution feature maps. We utilize a $7 \times 7 \times 7$ kernel and two $3 \times 3 \times 3$ kernels to enlarge the receptive field instead of stacking three widely used $3 \times 3 \times 3$. The next paragraph provides a detailed description of the overall model architecture.

The proposed network consists of a stem stage f^s, five residual stages $f^1, \ldots, f^5$, and a classification stage f^c. The positive and negative score vector, which is the output of proposed network, can be defined as $\hat{y} = f^c \circ f^5 \circ \ldots, \circ f^1 \circ f^s(\mathbf{x})$, where $\mathbf{x} \in [0, 1]^{220\times350\times160}$ is a preprocessed CT image. The stem stage f^s includes three convolution layers and a max-pooling layer, motivated by the work of [34]. All convolution layers in the proposed model are accompanied by batch normalization (BN) [31], and rectified linear unit (ReLU) activations [32]; these details have been excluded for the brevity of the paper. The first convolution layer of the stem stage f^s has $7 \times 7 \times 7$ kernels with a stride of 2. The other two convolution layers in the stem stage f^s have $3 \times 3 \times 3$ kernels with a stride of 1. The feature map extracted through three convolution layers outputs the pixels with the highest activation in each region through a max-pooling layer. In five residual stages $f^1, \ldots, f^5$, two types of bottleneck blocks [34] are employed: one with stride 1 and another with stride 2. The first residual stage f^1 has two bottleneck blocks with a stride of 1. The late four stages $f^2, \ldots, f^5$ have a bottleneck block with a stride of 1, and a bottleneck block with a stride of 2. The bottleneck block with a stride of 2 can be used in the late four stages $f^2, \ldots, f^5$ for down-sampling based on average-pooling. Each bottleneck block has three convolution layers with another convolution layer in skip connection. A bottleneck block with a stride of 2 includes an average-pooling layer in the main branch and skip connection. In the classification

stage f^c, the feature maps generated from five residual stages are compressed into (1, 1, 1) fixed size by global-average-pooling. Then, a fully connected (FC) layer outputs two scalar values from 512 scalar values. Finally, the positive and negative score vector $\hat{y}$ is output by a softmax layer. The source code for the proposed model is available at https://github.com/tkdgur658/GOTDNet.

IV. EXPERIMENTS

A. EXPERIMENTAL SETTINGS

The methods of Yao et al. [14] and Song et al. [11] were used as the comparative methods for the evaluation of the proposed method. Each method, including data preprocessing and DNN architecture, was reproduced based on our CT data. The experiments were implemented with Python 3.8 and PyTorch library 2.0. The weights of each model were optimized by using Focal loss [35] with a gamma value of 2.0. AdamW [36] was employed as the optimizer. The weight decay was set to 1e-4, and the learning rate varied from 1e-3 to 1e-6 by the cosine annealing scheduler over 50 epochs. Each model was optimized for a maximum of 50 epochs. An early stopping criterion was used, which terminated the training if there was no improvement in the loss for 20 epochs. The batch size was set to 8. In each experiment, the data was stratified and randomly sampled into three sets: 60% for training, 20% for validation, and 20% for testing. Training and testing were performed five times for each method. The training and testing were performed on two NVIDIA GeForce RTX 3090 GPUs with a data-parallel environment.

The diagnosis of treatment decisions can be considered a binary classification dealing with whether a treatment decision is recommended. Thus, we employed five measures: accuracy, F1 score, sensitivity, specificity, and Area under the receiver operating characteristics (AUC) curve, widely used in the study of binary classification to evaluate prediction performance. Descriptive statistics are presented as numbers for categorical variables and mean ± standard deviation for continuous variables. A paired t-test was performed to statistically evaluate the experimental results at the 0.1 and 0.05 significance levels. The null hypothesis states that the mean difference between the paired observations is zero. The t-test was performed for all possible pairs of comparison models using Scipy 1.5.4, which is an open-source Python library.

B. EXPERIMENTAL RESULTS

Table 2 summarizes the performance evaluations of the proposed method and two comparative methods. The experimental results show that the proposed method achieves the best performance across five metrics, rejecting the null hypothesis of statistical tests, which indicates the superiority of the proposed model. The proposed method achieved an average AUC of 0.793, outperforming the closest competitor by a substantial margin of 0.122. In terms of accuracy, the proposed model exhibited an average score of 0.699 with a standard deviation of 0.021, demonstrating an advantage of 0.072 over the second-best model. The F1 score followed a similar trend, with the proposed model leading at an average of 0.416 with a standard deviation of 0.018. This represented a difference of approximately 0.110 compared to the second-best model. Finally, the model achieved an average sensitivity of 0.723 and an average specificity of 0.694, which is considerably higher than all other models by a difference of 0.115 and 0.063, respectively.

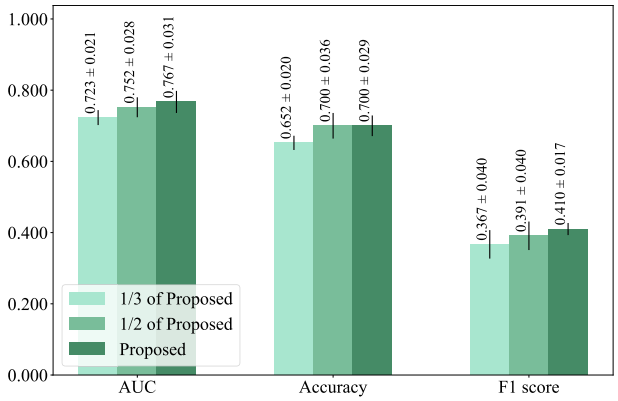


FIGURE 5. Performance changes over different input resolutions. Each bar represents the prediction performance when the CT data set is resized into different scales. For classification models, the model referenced as [11] and three different models in Fig. 4 were used to prevent rapid spatial information compression for each input scale. The brightest green indicates the performance when the smallest resolution input, at 1/3 the size recommended by the proposed preprocessing pipeline, was used with the model referenced as [11]. The medium bright green represents the performance using a medium resolution input, also at 1/3 the size suggested by the proposed preprocessing pipeline, with the model labeled 'a)' in Fig. 4. The dark green signifies the performance when an unresized dataset was input into the model referred to as 'b)' in Fig. 4.

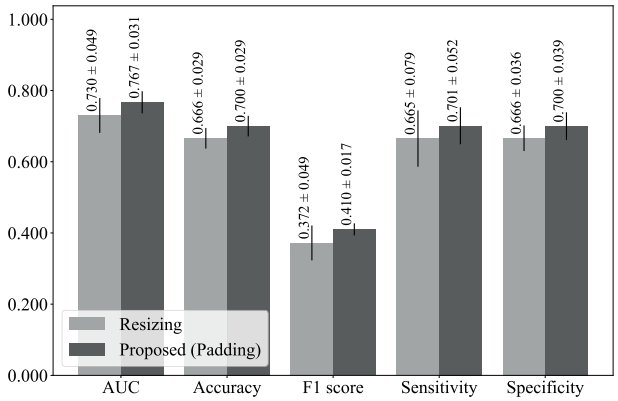


FIGURE 6. Performance improvements over different architecture modifications. The performance improvements over different modifications demonstrated in Fig. 4 are depicted. The proposed model achieves performance improvement through a four-stage structural enhancement from the baseline, which consists of adding a max-pooling and a residual stage, and using bottleneck block instead of basic block, and deep stem strategies. As a result, the average AUC improved from 0.728 to 0.793.

V. DISCUSSION

This study developed a DNN-based diagnosis method that can be applied to orbital CT images to make the treatment

TABLE 2. Experimental results. The average of the corresponding evaluation metric is presented in each cell with its standard deviation. The value in parentheses denotes the average ranking of the corresponding model. If the best scoring model for each metric rejects the null hypothesis in a *t*-test with all other comparison models, the asterisk (*) emphasizes the performance. * and ** indicate statistical significance at the 0.1 and 0.05 levels, respectively.

Method	AUC	Accuracy	F1 score	Sensitivity	Specificity
Proposed	**0.793 ± 0.022 (1.00)****	**0.699 ± 0.021 (1.00)****	**0.416 ± 0.018 (1.20)****	**0.723 ± 0.054 (1.20)***	**0.694 ± 0.024 (1.00)****
Yao et al. (2023)	0.671 ± 0.056 (2.40)	0.626 ± 0.049 (2.80)	0.326 ± 0.045 (2.40)	0.608 ± 0.096 (2.40)	0.630 ± 0.053 (2.80)
Song et al. (2021)	0.652 ± 0.066 (2.60)	0.627 ± 0.014 (2.20)	0.321 ± 0.064 (2.40)	0.599 ± 0.130 (2.40)	0.631 ± 0.024 (2.20)

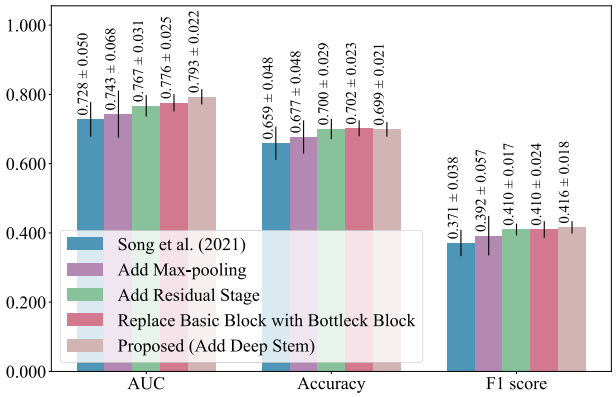


FIGURE 7. Performance changes over different deep stem strategies. The light, medium, and dark blue mean the basic stem strategy, the deep stem strategy with 3-3-3 kernels, and the deep stem strategy with 7-3-3 kernels, respectively.

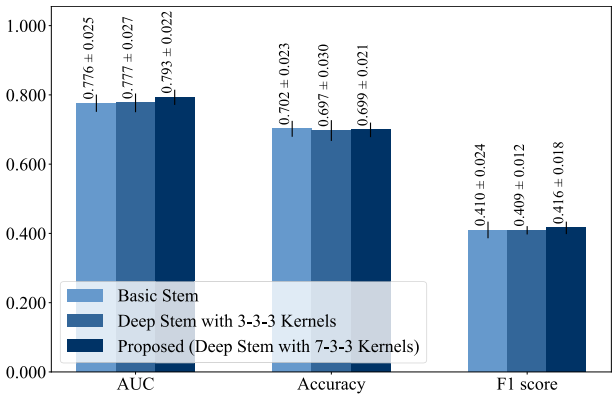


FIGURE 8. Performance changes over different architecture resizing scales. Each bar represents the prediction performance when our CT data set is resized to different scales. As classification models, three different models in Fig. 4 were used to prevent rapid spatial information compression for each data size. Light blue is the smallest input at 1/3 size Dark blue uses a dataset resized by 1/2. When resized to 1/2 size, a model with only max-pooling added is used. The bright green color indicates that no resizing dataset was input to the Modif. b model.

decisions of GO patients, achieving an average AUC of 0.793, accuracy of 0.699, F1 score of 0.416, sensitivity of 0.723, and specificity of 0.694. To the best of my knowledge, this is the first DNN-based GO patient treatment decision prediction study. In recent GO diagnosis studies, several DNN-based methods have been explored [12], [13], [23]. The activity of GO patients was evaluated using MRI [13] and CT images [23]. In addition, the severity of GO patients was classified using three different views of CT images [12]. Moreover, DNNs can be used for screening GO patients from normal cases [11]. However, studies related to GO treatment decisions have not yet been reported.

Several ablation studies provide evidence of the performance improvement achieved by the key designs of the proposed method (Fig. 5-8). In terms of data preprocessing, the proposed method includes a new pipeline compared to the previous two methods [11], [14]. Firstly, a pixel spacing of $0.35mm^3$ is used in the CT resampling process, while two existing methods use a pixel spacing of $1.00mm^3$. Several studies have been conducted in computer vision, focusing on performance improvement with high-resolution images [30]. Similarly, our study found that performance reduction can occur when we use low-resolution datasets demonstrated in Fig. 5. This observation can be further supported by the main experiments (Table 2), where the two comparison methods employing low-resolution resampling techniques consistently underperformed the proposed method. Second, the resizing approach of previous work [11] for providing a fixed input to the DNN can ignore spatial axis consistency; thus, we exploit zero-padding instead of resizing. As can be seen in Fig. 6, the padding approach yields an average AUC that is 0.037 higher than the resizing approach. Finally, a fine-grained orbital cropping strategy was employed to minimize training confusion from regions that are less relevant to GO, represented in Fig. 2.

In terms of DNN architecture, four modifications demonstrated in Fig. 4 significantly contribute to the performance improvement as depicted in Fig. 7. Firstly, the baseline model by Song et al. [11] exhibits the lowest performance due to drastic down-sampling in our dataset. The proposed model addresses this issue by incorporating max-pooling and an additional residual stage. The addition of max-pooling and residual stages can alleviate information loss from rapid down-sampling such as global average pooling. Specifically, the model of Song et al. [13], which was originally designed for low-resolution inputs (64, 128, 64), may not be well-suited for our high-resolution dataset with dimensions of (220, 350, 160).

Secondly, bottleneck blocks and average down-sampling can improve non-linearity and spatial information preservation. The basic residual blocks of the baseline use two $3 \times 3 \times 3$ kernels, which can make too many distortions in the spatial axis. Thus, we employ bottleneck blocks [34], which consist of two $1 \times 1 \times 1$ kernels and a $3 \times 3 \times 3$ kernel. Then, the convolution layer with stride 2 was replaced with average-pooling for low-resolution feature preservation.

These strategies can strengthen the nonlinearity of blocks while alleviating information loss in the spatial axis.

Finally, the deep stem strategy was exploited for enhancing high-resolution feature extraction. In the field of computer vision, deep stem layers are widely utilized to reduce computational costs in terms of the stem layer. However, the proposed model adopts the deep stem approach to expand the receptive field of low-level feature maps. The kernel size of the general deep stem layer is 3-3-3 [34], but the proposed model adopts a 7-3-3 kernel size for processing high-resolution images. As demonstrated in Fig. 8, the average AUC improves from 0.776 to 0.793, whereas the 3-3-3 deep stem reports 0.777.

Despite the notable results, this study has several limitations. First, the fine-grained cropping strategy in preprocessing is a labor intensive process. Thus, the automatic orbit cropping process should be considered for clinical applications in the future. Second, the proposed model has the potential for performance improvement by combining recent convolutional neural networks. Finally, a small number of training samples may limit performance improvement. In the future, label-efficient training methods should be considered to overcome the small number of training samples.

VI. CONCLUSION

This study introduced a DNN-based pipeline for making treatment decisions of Graves' orbitopathy patients using orbital CT images. The proposed method includes preprocessing and DNN, which highlights that exploiting the information from high-resolution CT images contributes to predictive performance. The proposed framework outperformed two existing methods, achieving superior results in AUC, accuracy, F1 score, sensitivity, and specificity. Moreover, our ablation study delineates the performance differences with existing models. Despite the promising outcomes, labor-intensive preprocessing, integration with advanced convolutional network architecture, and the limited number of training samples should be discussed in the future.

ACKNOWLEDGMENT

The authors would like to express their gratitude for the support of the KISTI Supercomputing Center, which facilitated the optimization of code performance for the experiments conducted in this study.

REFERENCES

[1] L. Bartalena and M. L. Tanda, "Current concepts regarding Graves' orbitopathy," *J. Internal Med.*, vol. 292, no. 5, pp. 692–716, Nov. 2022.

[2] C. Lei, M. Qu, H. Sun, J. Huang, J. Huang, X. Song, G. Zhai, and H. Zhou, "Facial expression of patients with Graves' orbitopathy," *J. Endocrinological Invest.*, vol. 46, no. 10, pp. 2055–2066, Apr. 2023.

[3] A.-M. Stancu, D. Alexandrescu, and C. Badiu, "Effects of block-replace regimen in patients with autoimmune hypothyroidism converted to Graves' disease," *Hormones*, vol. 23, no. 1, pp. 107–111, Mar. 2024.

[4] L. Bartalena and W. M. Wiersinga, "Proposal for standardization of primary and secondary outcomes in patients with active, moderate-to-severe Graves' orbitopathy," *Eur. Thyroid J.*, vol. 9, no. Suppl. 1, pp. 3–16, 2020.

[5] L. Bartalena, G. J. Kahaly, L. Baldeschi, C. M. Dayan, A. Eckstein, C. Marcocci, M. Marin, B. Vaidya, and W. M. Wiersinga, "The 2021 European group on graves' orbitopathy (EUGOGO) clinical practice guidelines for the medical management of graves' orbitopathy," *Eur. J. Endocrinol.*, vol. 185, no. 4, pp. G43–G67, 2021.

[6] Z. Li, Y. Luo, Q. Huang, Z. Chen, D. Song, D. Pan, S. Hu, W. Jiang, Q. Cai, X. Feng, Q. Zhang, C. Weng, Q. Zhong, T. Zhao, C. Li, T. Zhang, and J. Shen, "A randomized clinical trial of intravenous methylprednisolone with 2 protocols in patients with graves orbitopathy," *J. Clin. Endocrinology Metabolism*, vol. 109, no. 1, pp. 36–45, Dec. 2023.

[7] R. Luccas, C. M. Riguetto, M. Alves, D. E. Zantut-Wittmann, and F. Reis, "Computed tomography and magnetic resonance imaging approaches to Graves' ophthalmopathy: A narrative review," *Frontiers Endocrinology*, vol. 14, pp. 1–18, Jan. 2024.

[8] K. Arumugam, M. Naved, P. P. Shinde, O. Leiva-Chauca, A. Huaman-Osorio, and T. Gonzales-Yanac, "Multiple disease prediction using machine learning algorithms," *Mater. Today*, vol. 80, pp. 3682–3685, Jun. 2023.

[9] X. Liao, F. M. A. A. Aljufairi, K. K. H. Lai, K. K. W. Chan, R. Jia, W. Chen, Z. Hu, Y. Wei, W. C. W. Chu, C. C. Y. Tham, C. P. Pang, and K. K. L. Chong, "Clinical significance of corneal striae in thyroid associated orbitopathy," *J. Clin. Med.*, vol. 12, no. 6, p. 2284, Mar. 2023.

[10] J. Diao, X. Chen, Y. Shen, J. Li, Y. Chen, L. He, S. Chen, P. Mou, X. Ma, and R. Wei, "Research progress and application of artificial intelligence in thyroid associated ophthalmopathy," *Frontiers Cell Develop. Biol.*, vol. 11, pp. 1–18, Jan. 2023.

[11] X. Song, Z. Liu, L. Li, Z. Gao, X. Fan, G. Zhai, and H. Zhou, "Artificial intelligence CT screening model for thyroid-associated ophthalmopathy and tests under clinical conditions," *Int. J. Comput. Assist. Radiol. Surgery*, vol. 16, no. 2, pp. 323–330, Feb. 2021.

[12] J. Lee, W. Seo, J. Park, W.-S. Lim, J. Y. Oh, N. J. Moon, and J. K. Lee, "Neural network-based method for diagnosis and severity assessment of Graves' orbitopathy using orbital computed tomography," *Sci. Rep.*, vol. 12, no. 1, pp. 1–20, Jul. 2022.

[13] C. Lin, X. Song, L. Li, Y. Li, M. Jiang, R. Sun, H. Zhou, and X. Fan, "Detection of active and inactive phases of thyroid-associated ophthalmopathy using deep convolutional neural network," *BMC Ophthalmology*, vol. 21, no. 1, pp. 1–45, Dec. 2021.

[14] N. Yao, L. X. Li, Z. Y. Gao, C. Zhao, Y. T. Li, C. Han, J. F. Nan, Z. L. Zhu, Y. Xiao, F. B. Zhu, M. Zhao, and W. H. Zhou, "Deep learning-based diagnosis of disease activity in patients with graves' orbitopathy using orbital spect/Ct," *Eur. J. Nucl. Med. Mol. Imag.*, vol. 50, no. 12, pp. 3666–3674, Jun. 2023.

[15] X.-L. Bao, Y.-J. Sun, X. Zhan, and G.-Y. Li, "Orbital and eyelid diseases: The next breakthrough in artificial intelligence?" *Frontiers Cell Develop. Biol.*, vol. 10, pp. 4–48, Nov. 2022.

[16] R. Azad, A. Kazerouni, M. Heidari, E. K. Aghdam, A. Molaei, Y. Jia, A. Jose, R. Roy, and D. Merhof, "Advances in medical image analysis with vision transformers: A comprehensive review," *Med. Image Anal.*, vol. 91, Jan. 2024, Art. no. 103000.

[17] Y. Zeng, X. Zhang, Y. Kawasumi, A. Usui, K. Ichiji, M. Funayama, and N. Homma, "A 2.5D deep learning-based method for drowning diagnosis using post-mortem computed tomography," *IEEE J. Biomed. Health Informat.*, vol. 27, no. 2, pp. 1026–1035, Feb. 2023.

[18] S. Chang, Y. Gao, M. J. Pomeroy, T. Bai, H. Zhang, S. Lu, P. J. Pickhardt, A. Gupta, M. J. Reiter, E. S. Gould, and Z. Liang, "Exploring dual-energy CT spectral information for machine learning-driven lesion diagnosis in pre-log domain," *IEEE Trans. Med. Imag.*, vol. 42, no. 6, pp. 1835–1845, Sep. 2023.

[19] X. Li, R. Guo, J. Lu, T. Chen, and X. Qian, "Causality-driven graph neural network for early diagnosis of pancreatic cancer in non-contrast computerized tomography," *IEEE Trans. Med. Imag.*, vol. 42, no. 6, pp. 1656–1667, Jul. 2023.

[20] R. Kumar, P. Kumbharkar, S. Vanam, and S. Sharma, "Medical images classification using deep learning: A survey," *Multimedia Tools Appl.*, vol. 83, no. 7, pp. 19683–19728, Jul. 2023.

[21] S. H. Hosseini, R. Monsefi, and S. Shadroo, "Deep learning applications for lung cancer diagnosis: A systematic review," *Multimedia Tools Appl.*, vol. 83, no. 5, pp. 14305–14335, Jul. 2023.

[22] K. Hanai, H. Tabuchi, D. Nagasato, M. Tanabe, H. Masumoto, S. Miya, N. Nishio, H. Nakamura, and M. Hashimoto, "Automated detection of enlarged extraocular muscle in graves' ophthalmopathy with computed tomography and deep neural network," *Sci. Rep.*, vol. 12, no. 1, pp. 1–15, 2022.
[23] J. Lee, S. Lee, W. J. Lee, N. J. Moon, and J. K. Lee, "Neural network application for assessing thyroid-associated orbitopathy activity using orbital computed tomography," *Sci. Rep.*, vol. 13, no. 1, pp. 4–26, Aug. 2023.
[24] T. A. Soomro, L. Zheng, A. J. Afifi, A. Ali, S. Soomro, M. Yin, and J. Gao, "Image segmentation for MR brain tumor detection using machine learning: A review," *IEEE Rev. Biomed. Eng.*, vol. 16, pp. 70–90, 2023.
[25] S. Lee, J. K. Lee, and J. Lee, "Multihead neural network for multiple segmented images-based diagnosis of thyroid-associated orbitopathy activity," *IEEE Access*, vol. 12, pp. 43862–43873, 2024.
[26] R. Wang, T. Lei, R. Cui, B. Zhang, H. Meng, and A. K. Nandi, "Medical image segmentation using deep learning: A survey," *IET Image Process.*, vol. 16, no. 5, pp. 1243–1267, Apr. 2022.
[27] F. Zhu, Z. Gao, C. Zhao, Z. Zhu, J. Tang, Y. Liu, S. Tang, C. Jiang, X. Li, M. Zhao, and W. Zhou, "Semantic segmentation using deep learning to extract total extraocular muscles and optic nerve from orbital computed tomography images," *Optik*, vol. 244, Oct. 2021, Art. no. 167551.
[28] J. Hamwood, B. Schmutz, M. J. Collins, M. C. Allenby, and D. Alonso-Caneiro, "A deep learning method for automatic segmentation of the bony orbit in MRI and CT images," *Sci. Rep.*, vol. 11, no. 1, p. 26, Jul. 2021.
[29] L. Li, X. Song, Y. Guo, Y. Liu, R. Sun, H. Zou, H. Zhou, and X. Fan, "Deep convolutional neural networks for automatic detection of orbital blowout fractures," *J. Craniofacial Surgery*, vol. 31, no. 2, pp. 400–403, 2020.
[30] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.
[31] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.
[32] S. R. Dubey, S. K. Singh, and B. B. Chaudhuri, "Activation functions in deep learning: A comprehensive survey and benchmark," *Neurocomputing*, vol. 503, pp. 92–108, Sep. 2022.
[33] G. B. Bartley and C. A. Gorman, "Diagnostic criteria for Graves' ophthalmopathy," *Amer. J. Ophthalmology*, vol. 119, no. 6, pp. 792–795, Jun. 1995.
[34] H. Zhang, C. Wu, Z. Zhang, Y. Zhu, H. Lin, Z. Zhang, Y. Sun, T. He, J. Mueller, R. Manmatha, M. Li, and A. Smola, "ResNeSt: Split-attention networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 2735–2745.
[35] X. Zhong, G. Wang, W. Liu, Z. Wu, and Y. Deng, "Mask focal loss: A unifying framework for dense crowd counting with canonical object detection networks," *Multimedia Tools Appl.*, vol. 1, pp. 1–11, Jan. 2024.
[36] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *In: Proc. Int. Conf. Learn. Represent.*, 2019, pp. 1–16.

SANGHYUCK LEE is currently pursuing the Ph.D. degree in artificial intelligence with Chung-Ang University. He has published several papers applying machine learning algorithms to ophthalmology, including Graves' Orbitopathy. His research interests include computer vision and medical imaging.

MOHD ASYRAF ZULKIFLEY (Member, IEEE) received the bachelor's degree in engineering (mechatronics) from International Islamic University Malaysia, and the Ph.D. degree in electrical and electronics from The University of Melbourne. He completed a two-year Postdoctoral Fellowship in deep learning at the University of Oxford. His primary research interest includes deep learning for computer vision applications.

JEONG KYU LEE received the B.S. degree from Korea University College of Medicine, and the M.S. and Ph.D. degrees in ophthalmology from Korea University. He is currently an Ophthalmologist in oculoplastics, with expertise in eyelid and tear duct diseases, tumors, and thyroid eye diseases. He is also a Professor with the Department of Ophthalmology, Chung-Ang University College of Medicine, Chung-Ang University Hospital.

JAESUNG LEE received the B.S., M.S., and Ph.D. degrees in computer science from Chung-Ang University, Seoul, South Korea, in 2007, 2009, and 2013, respectively. He is currently an Associate Professor with the Department of Artificial Intelligence, Chung-Ang University, where he also works as the Chief of the AI/ML Innovation Research Center. His research interests include machine learning, multilabel learning, model selection, neural architecture search, feature selection, and multilabel learning with an emphasis on information theory.