scientific reports

Check for updates

OPEN

Deep learning model for tongue cancer diagnosis using endoscopic images

Jaesung Heo[1], June Hyuck Lim[1], Hye Ran Lee[2], Jeon Yeob Jang[2], Yoo Seob Shin[2], Dahee Kim[3], Jae Yol Lim[3], Young Min Park[3], Yoon Woo Koh[3], Soon-Hyun Ahn[4], Eun-Jae Chung[4], Doh Young Lee[4], Jungirl Seok[5] & Chul-Ho Kim[2]✉

In this study, we developed a deep learning model to identify patients with tongue cancer based on a validated dataset comprising oral endoscopic images. We retrospectively constructed a dataset of 12,400 verified endoscopic images from five university hospitals in South Korea, collected between 2010 and 2020 with the participation of otolaryngologists. To calculate the probability of malignancy using various convolutional neural network (CNN) architectures, several deep learning models were developed. Of the 12,400 total images, 5576 images related to the tongue were extracted. The CNN models showed a mean area under the receiver operating characteristic curve (AUROC) of 0.845 and a mean area under the precision-recall curve (AUPRC) of 0.892. The results indicate that the best model was DenseNet169 (AUROC 0.895 and AUPRC 0.918). The deep learning model, general physicians, and oncology specialists had sensitivities of 81.1%, 77.3%, and 91.7%; specificities of 86.8%, 75.0%, and 90.9%; and accuracies of 84.7%, 75.9%, and 91.2%, respectively. Meanwhile, fair agreement between the oncologist and the developed model was shown for cancer diagnosis (kappa value = 0.685). The deep learning model developed based on the verified endoscopic image dataset showed acceptable performance in tongue cancer diagnosis.

Oral cancer accounts for almost 3% of all cancer cases diagnosed worldwide[1]. According to the World Health Organization, more than 370,000 cases of oral cancer were reported in 2020[2]. Several studies have shown that tongue cancer is the most common type of oral cancer (42%)[3,4]. Oral cancer is prevalent in individuals mostly from Asia (65.8%) and is ranked one of Asia's sixth most frequent malignancies[5]. The lifestyle of the Asian population, which includes such as chain-smoking, alcohol consumption, and betel quid chewing, is a strong risk factor for oral cancer[6,7].

The early detection of tongue cancer is essential[8,9]. The overall 5-year survival rate for patients with tongue cancer is 68.1%[10]. According to the Surveillance, Epidemiology, and End Results database, the 5-year survival rates for local, regional, and distant stages are 82%, 68%, and 40%, respectively. In addition to the prognosis, patients with advanced tongue cancer experience difficulties during eating and speaking[11]. Furthermore, when the diagnosis is delayed, the scope of surgery broadens, and various invasive treatments are performed, resulting in increased side effects after treatment[12].

Endoscopy is a simple, effective, and non-invasive method for diagnosing tongue cancer[13]. However, only a few specialists have the ability to accurately read endoscopic results. For example, if a suspicious lesion is identified in a local clinic, the patient should be referred to a specialist for confirmation of disease status and further management[14]. However, general physicians who lack experience in treating patients with tongue cancer might mistakenly diagnose visual patterns for signs of ulceration or oral mucosa disease[15].

Studies on early detection of various malignancies using the characteristics of the tongue have been undertaken in the past[16–18]. Recently, the development of a primary diagnosis method through artificial intelligence (AI) analysis of oral endoscopic images can improve the chances of early diagnosis of tongue cancer. However, previous studies related to oral cancer were conducted with images created in non-clinical environments using

[1]Department of Radiation Oncology, Ajou University School of Medicine, Suwon, Republic of Korea. [2]Department of Otolaryngology, Ajou University School of Medicine, 164 Worldcup-ro, Yeongtong-gu, Suwon 16499, Republic of Korea. [3]Department of Otorhinolaryngology, Yonsei University, Seoul, Republic of Korea. [4]Department of Otorhinolaryngology-Head and Neck Surgery, Seoul National University Hospital, Seoul, Republic of Korea. [5]Department of Otorhinolaryngology-Head & Neck Surgery, National Cancer Center, Goyang, Republic of Korea. ✉email: ostium@ajou.ac.kr

Hospital	Diagnosis	n	%
AUH	Non-malignancy	1867	75.71
	Malignancy	599	24.29
SNUH	Non-malignancy	157	23.54
	Malignancy	510	76.46
NCC	Non-malignancy	648	94.74
	Malignancy	36	5.26
BRH	Non-malignancy	220	62.50
	Malignancy	132	37.50
YUH	Non-malignancy	743	52.81
	Malignancy	664	47.19
Total	Non-malignancy	3635	65.19
	Malignancy	1941	34.81
	Total	5576	100.00

Table 1. Dataset characteristics. *AUH* Ajou University Hospital, *SNUH* Seoul National University Hospital, *NCC* National Cancer Center, *BRH* Boramae Medical Center, *YUH* Yonsei University Hospital.

smartphones or digital cameras, rather than in a validated medical environment by using an endoscope; further, the number of images was small (< 300 images)[19,20]. In addition, studies have shown that there is a risk that the existing diagnostic algorithms may misdiagnose or underestimate the risk to critically ill patients in clinical applications[21]. This result was attributed to the low quality of the data collected for AI learning[22]. Hence, in this study, we verified the quality of the constructed dataset. Based on this data, we explored the feasibility of endoscopy-imaging-based deep learning models for tongue cancer diagnosis.

Results

Dataset characteristics. We retrospectively constructed the dataset of 12,400 verified endoscopic images obtained from five university hospitals in South Korea between 2010 and 2020. Of the 12,400 total images, 5576 images related to the tongue were extracted. For the development and validation of the total dataset (N = 5576), 1941 endoscopic images of malignant lesions and 3635 non-malignant endoscopic images were included. A difference in the ratio between malignant and non-malignant tumors was confirmed by each medical institution (Table 1). The internal validation dataset contained 1809 photographs of malignant lesions and 3415 non-malignant lesions. The external validation dataset consisted of 132 photographs of malignant lesions and 220 non-malignant lesions.

Parameter tuning and training. To perform fair comparison, all training hyperparameters were kept identical in all experiments (Fig. 1). The networks were trained for 300 epochs using binary cross-entropy loss with a batch size of 32. To avoid overfitting during training, we determined that overfitting occurred when the validation loss increased compared to the training loss, and then we explored ten additional epochs. If this trend continued, an early stopping logic that determines the parameter value in the epoch where the validation loss was increased compared to the train loss as the final parameter was applied. We did not use an algorithm that changes the learning rate according to the learning state, but rather applied Bayesian optimization to find the optimal learning rate to build the model.

Testing and model selection. After training, we evaluated the classification models using internal and external validation datasets. The evaluation results are summarized in Table 2. The optimal point of the ROC curve was determined when the AUROC reached its maximum value. When AUPRC, AUROC, specificity, and F1-score were compared for different models, DenseNet models showed excellent performance. Among them, Densenet169 had a higher AUROC, AUPRC, and accuracy than DenseNet 201 and DenseNet 121. Therefore, DensetNet169 was selected as the final model (Fig. 2).

AI vs. human readers. Figure 3 presents the test results for the best performing algorithm model and human readers on the external test dataset. The algorithm achieved an accuracy of 84.7% with a sensitivity of 81.1% and specificity of 86.8% for detecting tongue cancer. Among human readers, the accuracy of the oncology specialist was higher than that of the developed model at 92%. However, the accuracy of the general physician was lower than that of the model at 75.9%. The sensitivity and specificity considerably varied among the two human readers: the AI model achieved lower results than the specialist (sensitivity: 91.7%; specificity: 90.1%) and demonstrated significantly higher results than the general physician (sensitivity: 77.3%; specificity: 75.0%).

The agreement between the model and the human reader was estimated using the kappa value scale. Good agreement was observed between the model and the oncology specialist (kappa value = 0.685, 95% CI 0.606–0.763, $p < 0.001$). Further, moderate agreement was confirmed between the model and the general physician (kappa value = 0.482, 95% CI 0.389–0.575, $p < 0.001$). (Table 3).

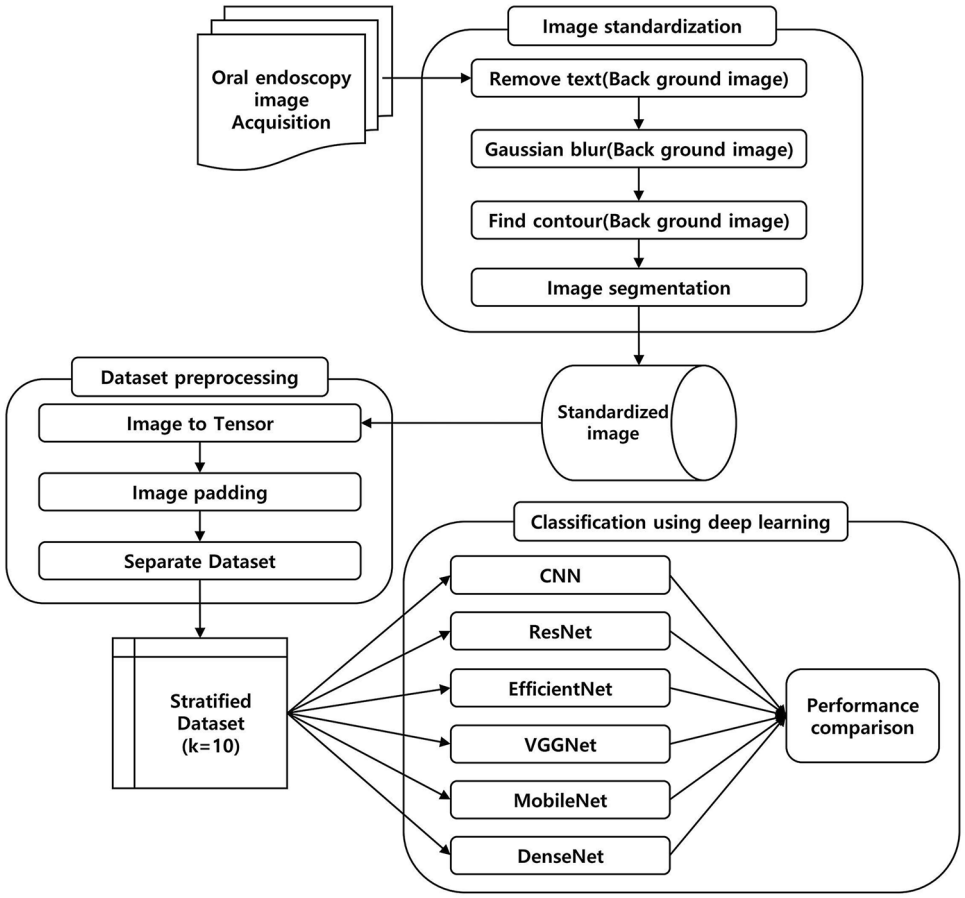

Figure 1. Overview of the development and evaluation of the tongue cancer diagnosis algorithm.

Discussion

This study developed a deep learning algorithm based on DenseNet169 with acceptable performance (i.e., AUROC 0.895 and AUPRC 0.918 for external validation datasets) for tongue cancer diagnosis from endoscopic images (Table 2 and Fig. 2). Other existing medical imaging studies have yielded higher results in some cases. However, unlike this study, most of them have a limitation in that they showed internal validation results rather than performance validation results when using an external test set[23,24]. The AI model developed in our study could derive the visual patterns of cancer in cluttered oral endoscopic images. This AI-based diagnostic tool could have clinical significance for the early diagnosis of cancer.

Although the diagnosis of tongue cancer should be made early, it is sometimes delayed[25]. In this case, as the cancer stage increases, the prognosis worsens, and the scope of surgery expands, resulting in severe postoperative side effects, such as dysarthria[26]. Early detection is difficult, and from the patient's viewpoint, knowledge and awareness regarding tongue cancer are lacking[27]. Furthermore, general physicians find it difficult to diagnose cancer in local areas using only endoscopic images[28]. Therefore, cancer should be diagnosed by an oncology specialist with extensive clinical experience. In previous studies, a screening system involving trained head and neck cancer specialists reduced oral cancer mortality[29].

However, the number of specialists is small, and most of them work in large medical institutions, including university hospitals with low accessibility to patients. In the present study, the developed deep learning model had superior performance in diagnosing cancer than a general physician, but inferior than an oncologist (Fig. 3). The difference between these results is possible because general physicians have relatively little clinical experience with cancer patients[30]. This indicates that AI-based diagnosis models have the potential to help general physicians with little clinical experience in oncology treatment to diagnose endoscopic images. In other studies, examples of increased cancer diagnostic accuracy have been provided with the aid of AI[31]. In addition, when considering the results of the kappa coefficient, there was a good agreement between the model developed in this study and the specialist in terms of lesion classification (kappa value = 0.685, 95% CI 0.606–0.763) (Table 3). Therefore, as in gastrointestinal endoscopy, the developed model will enable general physicians to improve the accuracy of diagnosing tongue cancer by combining it with oral endoscopy that is available in primary medical institutions.

Recently, several studies have reported the usefulness of medical image analysis based on deep learning models. The CNN model based on ResNet-50 simultaneously learned to detect and characterize lesions on magnetic resonance imaging (MRI)[32]. In addition, the developed CNN model with VGGNet classified benign or malignant lesions in medial image data[33]. In this study, we retrained an existing CNN model developed on a large general

Model	Sensitivity (95% CI)	Specificity (95% CI)	Precision (95% CI)	F1-score (95% CI)	Accuracy (95% CI)	AUROC (95% CI)	AUPRC (95% CI)
(a)							
CNN	0.712 (0.685–0.739)	0.860 (0.839–0.881)	0.733 (0.706–0.760)	0.720 (0.693–0.747)	0.809 (0.785–0.833)	0.882 (0.862–0.902)	0.932 (0.917–0.947)
VGG16	0.822 (0.799–0.845)	0.911 (0.894–0.928)	0.832 (0.809–0.855)	0.826 (0.803–0.849)	0.880 (0.860–0.900)	0.950 (0.937–0.963)	0.974 (0.964–0.984)
VGG19	0.801 (0.777–0.825)	0.910 (0.893–0.927)	0.828 (0.805–0.851)	0.813 (0.789–0.837)	0.872 (0.852–0.892)	0.941 (0.927–0.955)	0.969 (0.959–0.979)
DenseNet121	0.886 (0.867–0.905)	0.913 (0.896–0.930)	0.844 (0.822–0.866)	0.864 (0.843–0.885)	0.904 (0.886–0.922)	0.959 (0.947–0.971)	0.977 (0.968–0.986)
DenseNet169	0.890 (0.871–0.909)	0.921 (0.905–0.937)	0.859 (0.838–0.880)	0.873 (0.853–0.893)	0.910 (0.893–0.927)	0.960 (0.948–0.972)	0.977 (0.968–0.986)
DenseNet201	0.866 (0.845–0.887)	0.928 (0.912–0.944)	0.866 (0.845–0.887)	0.865 (0.844–0.886)	0.907 (0.889–0.925)	0.960 (0.948–0.972)	0.978 (0.969–0.987)
MobileNetV1	0.817 (0.794–0.840)	0.913 (0.896–0.930)	0.840 (0.818–0.862)	0.822 (0.799–0.845)	0.879 (0.859–0.899)	0.946 (0.932–0.960)	0.969 (0.959–0.979)
MobileNetV2	0.612 (0.582–0.642)	0.925 (0.909–0.941)	0.819 (0.796–0.842)	0.782 (0.757–0.807)	0.817 (0.794–0.840)	0.931 (0.916–0.946)	0.961 (0.949–0.973)
ResNet34	0.690 (0.662–0.718)	0.842 (0.820–0.864)	0.709 (0.681–0.737)	0.687 (0.659–0.715)	0.789 (0.764–0.814)	0.873 (0.853–0.893)	0.934 (0.919–0.949)
ResNet101	0.710 (0.683–0.737)	0.905 (0.887–0.923)	0.802 (0.778–0.826)	0.749 (0.723–0.775)	0.838 (0.816–0.860)	0.920 (0.904–0.936)	0.957 (0.945–0.969)
ResNet152	0.744 (0.718–0.770)	0.908 (0.891–0.925)	0.812 (0.788–0.836)	0.775 (0.750–0.800)	0.851 (0.829–0.873)	0.926 (0.910–0.942)	0.960 (0.948–0.972)
EfficientNetB3	0.618 (0.589–0.647)	0.920 (0.904–0.936)	0.804 (0.780–0.828)	0.681 (0.653–0.709)	0.815 (0.791–0.839)	0.899 (0.881–0.917)	0.944 (0.930–0.958)
(b)							
CNN	0.767 (0.723–0.811)	0.563 (0.511–0.615)	0.521 (0.469–0.573)	0.614 (0.563–0.665)	0.639 (0.589–0.689)	0.716 (0.669–0.763)	0.818 (0.778–0.858)
VGG16	0.701 (0.653–0.749)	0.821 (0.781–0.861)	0.706 (0.658–0.754)	0.700 (0.652–0.748)	0.776 (0.732–0.82)	0.866 (0.830–0.902)	0.917 (0.888–0.946)
VGG19	0.642 (0.592–0.692)	0.893 (0.861–0.925)	0.784 (0.741–0.827)	0.704 (0.656–0.752)	0.799 (0.757–0.841)	0.887 (0.854–0.920)	0.930 (0.903–0.957)
DenseNet121	0.795 (0.753–0.837)	0.831 (0.792–0.870)	0.750 (0.705–0.795)	0.765 (0.721–0.809)	0.817 (0.777–0.857)	0.885 (0.852–0.918)	0.906 (0.876–0.936)
DenseNet169	0.793 (0.751–0.835)	0.853 (0.816–0.890)	0.773 (0.729–0.817)	0.777 (0.734–0.82)	0.830 (0.791–0.869)	0.895 (0.863–0.927)	0.918 (0.889–0.947)
DenseNet201	0.769 (0.725–0.813)	0.876 (0.842–0.910)	0.793 (0.751–0.835)	0.778 (0.735–0.821)	0.836 (0.797–0.875)	0.892 (0.860–0.924)	0.913 (0.884–0.942)
MobileNetV1	0.701 (0.653–0.749)	0.878 (0.844–0.912)	0.789 (0.746–0.832)	0.730 (0.684–0.776)	0.811 (0.77–0.852)	0.884 (0.851–0.917)	0.906 (0.876–0.936)
MobileNetV2	0.435 (0.383–0.487)	0.909 (0.879–0.939)	0.757 (0.712–0.802)	0.619 (0.568–0.67)	0.732 (0.686–0.778)	0.802 (0.760–0.844)	0.847 (0.809–0.885)
ResNet34	0.674 (0.625–0.723)	0.717 (0.670–0.764)	0.607 (0.556–0.658)	0.623 (0.572–0.674)	0.701 (0.653–0.749)	0.793 (0.751–0.835)	0.871 (0.836–0.906)
ResNet101	0.532 (0.480–0.584)	0.883 (0.849–0.917)	0.741 (0.695–0.787)	0.612 (0.561–0.663)	0.751 (0.706–0.796)	0.842 (0.804–0.880)	0.902 (0.871–0.933)
ResNet152	0.662 (0.613–0.711)	0.856 (0.819–0.893)	0.744 (0.698–0.79)	0.695 (0.647–0.743)	0.783 (0.740–0.826)	0.856 (0.819–0.893)	0.908 (0.878–0.938)
EfficientNetB3	0.524 (0.472–0.576)	0.865 (0.829–0.901)	0.739 (0.693–0.785)	0.572 (0.520–0.624)	0.737 (0.691–0.783)	0.816 (0.776–0.856)	0.873 (0.838–0.908)

Table 2. Diagnostic performance of CNN models in internal validation (a) and external validation (b).

natural image dataset using oral endoscopic images (Fig. 1). Six different models were used in this study: CNN, ResNet, EfficientNet, VGGNet, MobileNet, and DenseNet. Because the CNN model is the most basic model for image classification, it was used as a basis for comparing the performances of other models.

VGGNet, ResNet, and DenseNet were models that share a huge skeleton, and when the layers are deepened, each model can achieve better prediction performance. We were able to spot trends in the data and find an appropriate model using these associated models. MobileNet and VGGNet have relatively fast learning speeds, comply with the required performance, and are used to quickly check the results by adding logic to find data features more efficiently. ResNet, DenseNet, and EfficientNet are composed of deep layers; therefore, their learning speed is relatively slow, but their performance is acceptable. In particular, DenseNet shows superior performance with fewer parameters than ResNet. ResNet combines features by summation when passing through layers, but DenseNet is different because it concatenates the features rather than adding them.

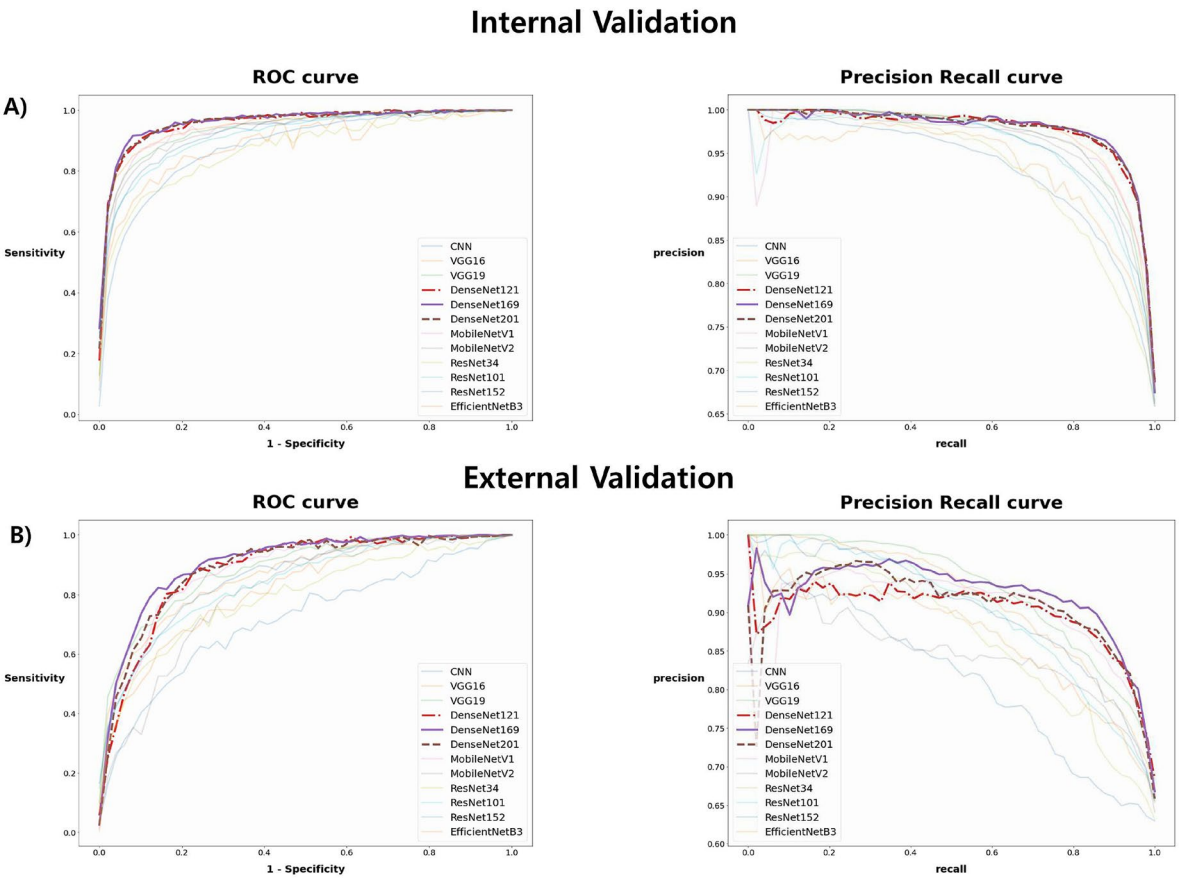

Figure 2. Receiver operating characteristic curves and precision-recall curves for the deep learning algorithm on internal validation dataset (**A**) and external validation datasets (**B**).

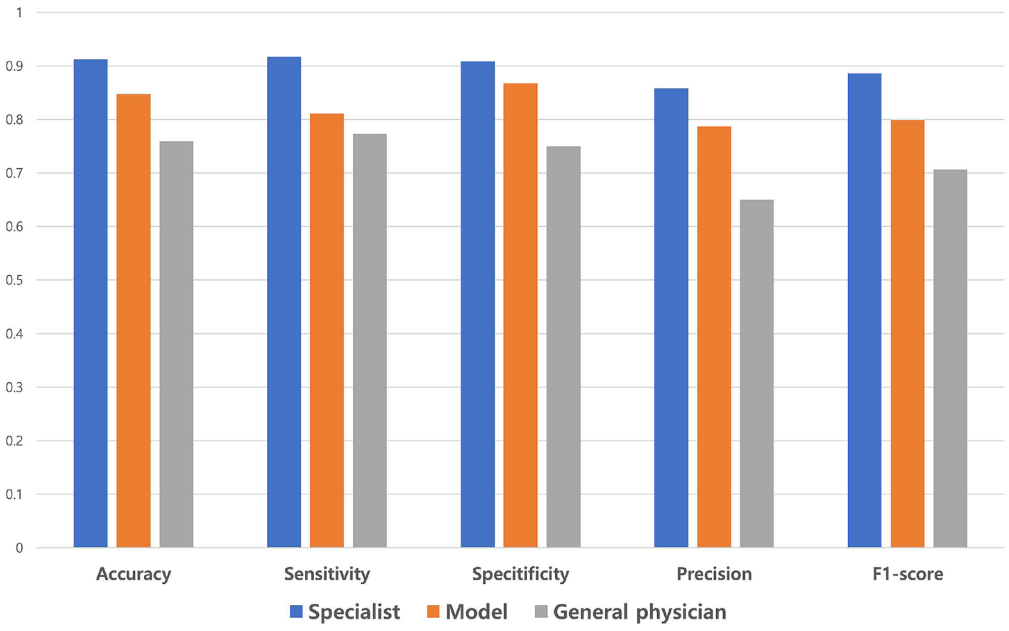

Figure 3. Performance of the deep learning model and comparison with human readers.

	Malignancy prediction		
	Kappa value	95% CI	P value
Model vs			
Specialist	0.685	0.606–0.763	< 0.001
General physician	0.482	0.389–0.575	< 0.001

Table 3. Agreement of the model and human readers.

Unlike previous studies that used standardized CT and MRI images, this study analyzed atypical oral images using the deep learning algorithm mentioned above. Since tongue cancer is a rare disease, we removed as much noise as possible from the image rather than increasing the amount of data. By minimizing the deviation of the data, the difference between the sample population and the overall population was reduced. DenseNet169, which was evaluated as the most suitable algorithm in this study, was also effective in image evaluation conducted in previous studies. In a study to classify pathological images in which atypical images were used similar to this study, effective results were obtained even with a small number of images[34]. Similarly, DenseNet169 showed the best performance in the study of the AI model for classifying the quality of tongue images[35]. Therefore, the application and optimization of AI algorithms considering the characteristics of each image data is essential. In particular, we believe that the model derived from this study will be meaningful for atypical data with large deviations among images, including endoscopes.

Despite recent innovative advances in deep learning technology, a large, validated dataset is one of the prerequisites for improving diagnostic performance. Driks emphasized the problem of "Frankenstein datasets"[22]. A Frankenstein dataset comprises information collected from multiple sources and assembled piece by piece. If an algorithm is tested with the same data used to train the model, it tends to appear to perform more accurately than it actually would on more realistic data or in practical applications. Therefore, we focused on well-organized and high-quality dataset construction. In the previous study, easily accessible smartphone and digital camera images were used; however, in this study, a dataset was constructed using oral endoscopy images created in clinical sites[19,20]. The poor-quality images could affect the analysis of image features and directly lead to a wrong diagnosis, causing severe interference with the development of the AI model. Therefore, oral endoscopic images are difficult to classify. In particular, oral endoscopy performed during the treatment process has different characteristics depending on the examiner because no guidelines were set for imaging.

This medical condition could lead to incorporation bias in the dataset. To create a relatively stable tongue image dataset, tongue images were collected using uniform endoscopic equipment. Additionally, to improve the quality of the dataset, several head and neck cancer specialists from multiple institutions directly participated in the data collection and review process. De-identification of data was carried out, and data inspection was performed more than twice. Moreover, a verification was conducted by TTA, an external institution. The radiomics approach used in previous studies involves manual ROI segmentation and extraction of several text features[36]. However, in this study, a deep learning network can be trained automatically without ROI segmentation. Therefore, advantages exist in terms of the decreased training time and costs for annotation workers. This method is designed to extract features directly from a dataset without the prerequisite for segmentation and manual processing. We performed processing to remove areas other than important areas so that the model could easily identify patterns in the image data.

We preprocessed the dataset before developing the AI model. The endoscopic images were of varying sizes, lighting conditions, and angles. In addition, owing to the noise of the equipment itself, some pixels sporadically entered as outliers in the oral endoscopy image. Some images also contained textual information, such as weather, and provided line guidelines (Fig. 1). In addition to the previous data preprocessing steps, such as scaling and adjusting the exposure, we developed a new algorithm and applied it in our research. For image standardization, we proceeded as follows. (1) We created a background image by converting the target image into a black-and-white image. (2) We removed the text from the background images. (3) We blurred the background image based on outliers using Gaussian blur. (4) The lesions were explored in the background image. (5) We cropped useless parts from the original image based on the lesions found in the background image (Supplement 1). All images were then converted into the JPEG format as required by our deep neural framework. According to the algorithms, they were then resized to 224×224 or 300×300 to the required input image size of the models before the model training process.

The current study had several limitations. First, the developed model cannot make a definite diagnosis for benign diseases among tongue lesions, such as leukoplakia and ulcers. In future studies, we plan to develop a model that can clearly distinguish benign and malignant tumors by classifying them into three categories: normal, benign, and malignant. Second, the oral endoscopic image characteristics used in this study differed from those of conventional CT and MRI images. These data have a high degree of freedom and are affected by the features of the endoscope user with atypical, non-standardized images. We used various data preprocessing techniques to compensate for these shortcomings. When collecting data in future studies, it would be beneficial to consider the application of endoscopy guidelines. Third, developing a cancer diagnosis model using only endoscopic images has a limit. In future research, high-performance diagnostic models are expected to be developed if images are combined with various clinical data. Fourth, several medical institutions participated in this study, resulting in differences between institutions in the amount of data, image characteristics, and the ratio of malignancy to non-malignancy (Table 1). In this study, data preprocessing was performed to correct this. In future research,

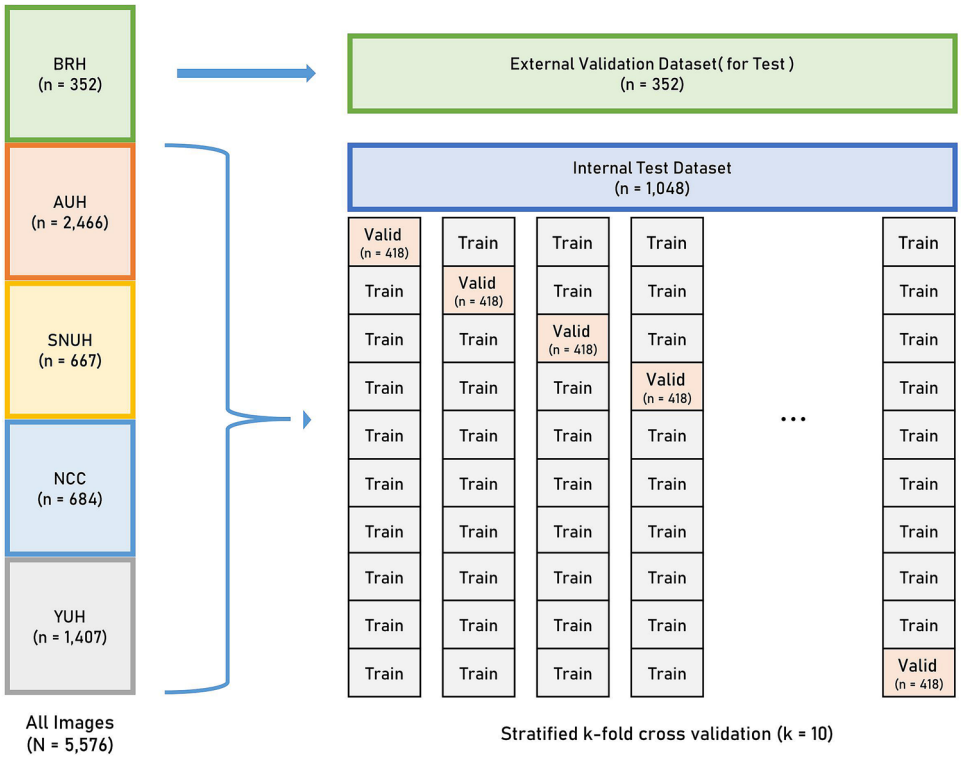

Figure 4. Validation and test structure diagram of the tongue cancer dataset for deep learning.

uniformly distributing the ratio and amount of data for each participating institution would be necessary. Finally, lesions were not detected in this study. In future work, we plan to collect additional information on lesions and use it to develop an AI model that identifies suspected lesions with heat maps using Grad-CAM.

In conclusion, we have constructed a quality-validated dataset using oral endoscopy images from several medical institutions. A deep learning model based on the dataset showed acceptable performance for application in tongue cancer diagnosis. Compared with human readers, it showed lower diagnostic performance than oncology specialists and higher diagnostic performance than general physicians. Therefore, the developed algorithm could be used as an assistant tool for general physicians to increase the diagnosis and screening of cancer in clinical settings.

Methods

Dataset. We retrospectively collected 12,400 clinical endoscopic images from five hospitals in South Korea (i.e., Seoul National University Hospital, Yonsei University Hospital, Ajou University Hospital, National Cancer Center, and Boramae Medical Center) between December 9, 2010, and September 24, 2020. Through a database query of the medical databases (i.e., EMR and PACS), we extracted the endoscopic images taken for diagnosis of tongue cancer and the pathological reports of the images. The extracted endoscopic images were read and reviewed by at least two head and neck oncologists at each hospital, and image preprocessing, such as de-identification, was performed. The diagnosis results of each oral imaging image can be classified as malignant, benign, or normal. Among these, benign and normal images were classified as non-malignant images. The constructed dataset has undergone and passed an external verification by the Telecommunications Technology Association (TTA) for data structure and format accuracy.

Of the 5576 total tongue images, we selected 5224 images (internal validation dataset) to develop the algorithm and then used the remaining 352 images (external validation dataset) for testing (Fig. 4). Pathological diagnosis was used as the correct answer to develop and validate the deep learning model. The Institutional Review Board of Ajou University Hospital approved this study (IRB No. AJIRB-MDB-20-311). Further, informed consent from all participants was waived by the IRB because of the retrospective nature of this study. All methods were performed in accordance with the Declaration of Helsinki.

Deep learning model. To detect malignancy from oral endoscopic images (Fig. 1), we developed an automated deep learning algorithm using a cascaded convolutional neural network (CNN). The backbone networks for the detection and classification were initialized using a pre-trained model, which was trained with tens of millions of images in the ImageNet dataset and was further finetuned using the development dataset[37]. The tensor converted from the image was subjected to data scaling, data-type adjustment, and padding to maintain the image ratio. To optimize the hyperparameter, we used a Bayesian optimization method for the training and

internal validation processes[38]. The target of Bayesian optimization was the area under a receiver operating characteristic curve (AUROC), and the hyperparameters that maximize AUROC were derived. The minibatch size was determined to be 32 to further improve the generalization performance. After the optimal hyperparameters were determined, we obtained the best model and evaluated its performance in the testing set.

A CNN architecture was constructed to calculate the probability of malignancy of an endoscopic image using ResNet (i.e., ResNet34, ResNet101, and ResNet152)[39], EfficientNet B3[40], VGGNet (i.e., VGG 16 and VGG 19)[41], MobileNet (i.e., MobileNetV1 and MobileNetV2)[42], and DenseNet (i.e., DenseNet121, DenseNet169, and DenseNet201)[43]. These models are neural networks with several layers and are commonly used for image classification. We applied stratified k-fold cross-validation to assess the deep learning model (k = 10). A total of 10 random datasets were extracted by fixing the seeds to ensure that the non-malignant and malignant ratios were equal. During internal validation, we randomly partitioned the dataset into approximately 70% training, 10% validation, and 20% test sets (Fig. 4). Moreover, we determined the number of epochs using an early termination tool. In this process, a dataset consisting of images obtained from Seoul National University Hospital, Severance Hospital, Ajou University Hospital, and National Cancer Center was used for internal validation. Through this method, the risk of overfitting increases from the moment the validation loss increases compared to the training loss. Thus, the training was ended after additional exploration.

After training the models, we examined the accuracy of the trained models by other clinical research centers in distinguishing non-malignant from malignant for external validation. To this end, we constructed a new testing dataset including 352 tongue images using the Boramae Medical Center dataset.

Comparison with observer classification. We compared the performance of the algorithm with that of human readers using an external validation dataset. The human readers employed in our study were divided into two groups according to their professional backgrounds and clinical experiences. A specialist human reader who was a head and neck surgical oncologist with more than seven years of clinical experience participated in this study. The general physician human reader was a doctor with four years of experience after obtaining their license and was a non-specialist.

The human reader reviewed the same dataset and classified cases as malignant vs. non-malignant, without any prior knowledge on the patient history. The reader blindly evaluated the de-identified endoscopic image of the data and assessed the possibility of malignancy. The AI model with the best performance among the models was also evaluated using the same dataset.

The performance of the readers was assessed by comparing their predictions with the corresponding pathological reports. We evaluated the final results and calculated the overall accuracy, sensitivity, and specificity. We estimated the kappa values with linear weighting and 95% confidence intervals (CIs) to compare the diagnostic results of human readers and the model. The kappa value scale for agreement strength was as follows: poor: < 0.2; fair: 0.21–0.40; moderate: 0.41–0.60; good: 0.61–0.80; and very good: 0.81–1.00[44].

Statistical analysis. We evaluated the performance of the classification models using objective evaluation metrics, including specificity, precision, sensitivity, F1-score, and accuracy. The metrics base their mathematical foundation on the true positive (TP), true negative, false negative, and false-positive (FP) values of the models' predictions. In addition, we used AUROC to evaluate the performance of the deep learning algorithm for distinguishing malignant from non-malignant. We plotted the receiver operating characteristic (ROC) curve by calculating the TP rate (sensitivity) and the FP rate (1 − specificity) with different predicted probability thresholds, and then we calculated the AUC values. Because the distribution of binary cases was not uniform, we also estimated the area under the precision-recall curve (AUPRC) values to evaluate the trained model. The corresponding 95% confidence interval was computed for each indicator value. The performance of the CNN models and the two readers in distinguishing malignant from non-malignant images was evaluated using these indicators.

We selected the model that best classified the endoscopic images by comparing the model performance. When selecting a model, the performance was evaluated by considering the first AUROC and the second AUPRC. Even if the model showed high performance in internal validation, the model that showed poor performance for external validation was excluded from model selection. All statistical analyses were performed using pandas (version 0.22.1), scikit-learn (version 0.24.1), NumPy (1.19.5), Matplotlib (3.3.4), OpenCV-Python (4.5.2), and Bayesian optimization (1.2.0) Python packages. We used Keras, which is a deep learning framework that acts as an interface for the TensorFlow2 library. Model structures were developed on graphical processing unit servers with multiple NVIDIA Tesla V100 graphic process units (32 GB × 4) and Xeon Gold 6248 (2.5 GHZ/20-core/150 W, 512 GB RAM) as the central processing unit.

Ethical statement. The Institutional Review Board of Ajou University Hospital approved this study (IRB No. AJIRB-MDB-20–311). Further, informed consent from all participants was waived by the IRB because of the retrospective nature of this study.

Data availability

The datasets generated and/or analyzed in this study are available from the corresponding author upon reasonable request.

Code availability

To train the classification model in this study, we used the publicly available TensorFlow training script available at https://github.com/tensorflow/models/tree/master/official/vision/image_classification.

Received: 28 September 2021; Accepted: 29 March 2022
Published online: 15 April 2022

References

1. Bray, F. *et al.* Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **68**, 394–424. https://doi.org/10.3322/caac.21492 (2018).
2. Ren, Z.-H., Hu, C.-Y., He, H.-R., Li, Y.-J. & Lyu, J. Global and regional burdens of oral cancer from 1990 to 2017: Results from the global burden of disease study. *Cancer Commun.* **40**, 81–92. https://doi.org/10.1002/cac2.12009 (2020).
3. Razmpa, E., Memari, F. & Naghibzadeh, B. Epidemiologic and clinicopathologic characteristics of tongue cancer in Iranian patients. *Acta Med. Iran.* **49**, 44–48 (2011).
4. Ibayashi, H. *et al.* Estimation of premature mortality from oral cancer in Japan, 1995 and 2005. *Cancer Epidemiol.* **35**, 342–344. https://doi.org/10.1016/j.canep.2011.01.010 (2011).
5. Krishna Rao, S. V., Mejia, G., Roberts-Thomson, K. & Logan, R. Epidemiology of oral cancer in Asia in the past decade—An update (2000–2012). *Asian Pac. J. Cancer Prev.* **14**, 5567–5577. https://doi.org/10.7314/apjcp.2013.14.10.5567 (2013).
6. Hashibe, M. *et al.* Alcohol drinking in never users of tobacco, cigarette smoking in never drinkers, and the risk of head and neck cancer: Pooled analysis in the International Head and Neck Cancer Epidemiology Consortium. *J. Natl. Cancer Inst.* **99**, 777–789. https://doi.org/10.1093/jnci/djk179 (2007).
7. Guha, N., Warnakulasuriya, S., Vlaanderen, J. & Straif, K. Betel quid chewing and the risk of oral and oropharyngeal cancers: A meta-analysis with implications for cancer control. *Int. J. Cancer* **135**, 1433–1443. https://doi.org/10.1002/ijc.28643 (2014).
8. Subramanian, S. *et al.* Cost-effectiveness of oral cancer screening: Results from a cluster randomized controlled trial in India. *Bull. World Health Organ.* **87**, 200–206. https://doi.org/10.2471/blt.08.053231 (2009).
9. Awan, K. Oral cancer: Early detection is crucial. *J. Int. Oral. Health* **6**, i–ii (2014).
10. SEER Cancer Stat Facts: Tongue Cancer. National Cancer Institute. Bethesda, MD. https://seer.cancer.gov/statfacts/html/tongue.html. (Accessed 2 September 2021).
11. Philiponis, G. & Kagan, S. H. Speaking legibly: Qualitative perceptions of altered voice among oral tongue cancer survivors. *Asia Pac. J. Oncol. Nurs.* **2**, 250–256. https://doi.org/10.4103/2347-5625.158020 (2015).
12. de Melo, G. M., Ribeiro, K. D. C. B., Kowalski, L. P. & Deheinzelin, D. Risk factors for postoperative complications in oral cancer and their prognostic implications. *Arch. Otolaryngol. Head Neck Surg.* **127**, 828–833 (2001).
13. Thong, P. S. P. *et al.* Clinical application of fluorescence endoscopic imaging using hypericin for the diagnosis of human oral cavity lesions. *Br. J. Cancer* **101**, 1580–1584. https://doi.org/10.1038/sj.bjc.6605357 (2009).
14. Grafton-Clarke, C., Chen, K. W. & Wilcock, J. Diagnosis and referral delays in primary care for oral squamous cell cancer: A systematic review. *Br. J. Gen. Pract.* **69**, e112–e126. https://doi.org/10.3399/bjgp18X700205 (2019).
15. Jafari, A., Najafi, S., Moradi, F., Kharazifard, M. & Khami, M. Delay in the diagnosis and treatment of oral cancer. *J. Dent. (Shiraz)* **14**, 146–150 (2013).
16. Han, S. *et al.* Potential screening and early diagnosis method for cancer: Tongue diagnosis. *Int. J. Oncol.* **48**, 2257–2264. https://doi.org/10.3892/ijo.2016.3466 (2016).
17. Lo, L. C., Cheng, T. L., Chen, Y. J., Natsagdorj, S. & Chiang, J. Y. TCM tongue diagnosis index of early-stage breast cancer. *Complement Ther. Med.* **23**, 705–713. https://doi.org/10.1016/j.ctim.2015.07.001 (2015).
18. Song, A. Y. *et al.* Diagnosis of early esophageal cancer based on TCM tongue inspection. *Biomed. Environ. Sci.* **33**, 718–722. https://doi.org/10.3967/bes2020.094 (2020).
19. Fu, Q. *et al.* A deep learning algorithm for detection of oral cavity squamous cell carcinoma from photographic images: A retrospective study. *EClinicalMedicine* **27**, 100558. https://doi.org/10.1016/j.eclinm.2020.100558 (2020).
20. Song, B. *et al.* Automatic classification of dual-modalilty, smartphone-based oral dysplasia and malignancy images using deep learning. *Biomed. Opt. Express* **9**, 5318–5329. https://doi.org/10.1364/BOE.9.005318 (2018).
21. Wynants, L. *et al.* Prediction models for diagnosis and prognosis of covid-19: Systematic review and critical appraisal. *BMJ* **369**, m1328. https://doi.org/10.1136/bmj.m1328 (2020).
22. Roberts, M. *et al.* Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nat. Mach. Intell.* **3**, 199–217. https://doi.org/10.1038/s42256-021-00307-0 (2021).
23. Jang, B.-S. *et al.* Image-based deep learning model for predicting pathological response in rectal cancer using post-chemoradiotherapy magnetic resonance imaging. *Radiother. Oncol.* **161**, 183–190. https://doi.org/10.1016/j.radonc.2021.06.019 (2021).
24. Horvat, N. *et al.* MR imaging of rectal cancer: Radiomics analysis to assess treatment response after neoadjuvant therapy. *Radiology* **287**, 833–843. https://doi.org/10.1148/radiol.2018172300 (2018).
25. Marella, G. L. *et al.* The diagnostic delay of oral carcinoma. *Ig Sanita Pubbl* **74**, 249–263 (2018).
26. Hutcheson, K. A. & Lewin, J. S. Functional assessment and rehabilitation: How to maximize outcomes. *Otolaryngol. Clin. N. Am.* **46**, 657–670. https://doi.org/10.1016/j.otc.2013.04.006 (2013).
27. Warnakulasuriya, K. A. *et al.* An alarming lack of public awareness towards oral cancer. *Br. Dent. J.* **187**, 319–322. https://doi.org/10.1038/sj.bdj.4800269 (1999).
28. Sargeran, K., Murtomaa, H., Safavi, S. M. & Teronen, O. Delayed diagnosis of oral cancer in Iran: Challenge for prevention. *Oral. Health Prev. Dent.* **7**, 69–76 (2009).
29. Chuang, S. L. *et al.* Population-based screening program for reducing oral cancer mortality in 2,334,299 Taiwanese cigarette smokers and/or betel quid chewers. *Cancer* **123**, 1597–1609. https://doi.org/10.1002/cncr.30517 (2017).
30. Round, T., Steed, L., Shankleman, J., Bourke, L. & Risi, L. Primary care delays in diagnosing cancer: What is causing them and what can we do about them?. *J. R. Soc. Med.* **106**, 437–440. https://doi.org/10.1177/0141076813504744 (2013).
31. Huang, S., Yang, J., Fong, S. & Zhao, Q. Artificial intelligence in cancer diagnosis and prognosis: Opportunities and challenges. *Cancer Lett.* **471**, 61–71. https://doi.org/10.1016/j.canlet.2019.12.007 (2020).
32. Herent, P. *et al.* Detection and characterization of MRI breast lesions using deep learning. *Diagn. Interv. Imaging* **100**, 219–225. https://doi.org/10.1016/j.diii.2019.02.008 (2019).
33. Antropova, N., Abe, H. & Giger, M. L. Use of clinical MRI maximum intensity projections for improved breast lesion classification with deep convolutional neural networks. *J. Med. Imaging (Bellingham)* **5**, 014503–014503. https://doi.org/10.1117/1.JMI.5.1.014503 (2018).
34. Mohammed, M. A., Abdurahman, F. & Ayalew, Y. A. Single-cell conventional pap smear image classification using pre-trained deep neural network architectures. *BMC Biomed. Eng.* **3**, 11. https://doi.org/10.1186/s42490-021-00056-6 (2021).
35. Jiang, T. *et al.* Tongue image quality assessment based on a deep convolutional neural network. *BMC Med. Inform. Decis. Mak.* **21**, 147. https://doi.org/10.1186/s12911-021-01508-8 (2021).
36. van Timmeren, J. E., Cester, D., Tanadini-Lang, S., Alkadhi, H. & Baessler, B. Radiomics in medical imaging—"how-to" guide and critical reflection. *Insights Imaging* **11**, 91. https://doi.org/10.1186/s13244-020-00887-2 (2020).
37. Deng, J. *et al.* In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 248–255.
38. Wu, J. *et al.* Hyperparameter optimization for machine learning models based on Bayesian optimizationb. *J. Electron. Sci. Technol.* **17**, 26–40. https://doi.org/10.11989/JEST.1674-862X.80904120 (2019).

39. Guo, S. & Yang, Z. Multi-Channel-ResNet: An integration framework towards skin lesion analysis. *Inform. Med. Unlocked* **12**, 67–74. https://doi.org/10.1016/j.imu.2018.06.006 (2018).
40. Alzubaidi, L. *et al.* Novel transfer learning approach for medical imaging with limited labeled data. *Cancers (Basel)* **13**, 1590. https://doi.org/10.3390/cancers13071590 (2021).
41. Hesamian, M. H., Jia, W., He, X. & Kennedy, P. Deep learning techniques for medical image segmentation: Achievements and challenges. *J. Digit. Imaging* **32**, 582–596. https://doi.org/10.1007/s10278-019-00227-x (2019).
42. Wang, W. *et al.* A new image classification approach via improved MobileNet models with local receptive field expansion in shallow layers. *Comput. Intell. Neurosci.* **2020**, 8817849. https://doi.org/10.1155/2020/8817849 (2020).
43. Gottapu, R. D. & Dagli, C. H. DenseNet for anatomical brain segmentation. *Proc. Comput. Sci.* **140**, 179–185. https://doi.org/10.1016/j.procs.2018.10.327 (2018).
44. Landis, J. R. & Koch, G. G. The measurement of observer agreement for categorical data. *Biometrics* **33**, 159–174. https://doi.org/10.2307/2529310 (1977).

Acknowledgements
We would like to thank Editage (www.editage.co.kr) for editing and reviewing this manuscript for English language.

Author contributions
Conceived and designed the analysis: C.K., J.H. Collected the data: H.R.L., J.Y.J., Y.S.S., D.K., J.Y.L., Y.M.P., Y.W.K., S.H.A., E.J.C., D.Y.L., J.S. Contributed data or analysis tools: J.H., J.H.L. Performed the analysis: J.H. Wrote the paper: J.H., J.H.L. Manuscript editing: C.K.

Funding
This research was supported by the Bio & Medical Technology Development Program of the National Research Foundation (NRF) funded by the Korean government (MSIT) (No. 2021R1G1A1095212, 2017M3A9F7079339). This research was supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (Grant number: HR21C1003).

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-10287-9.

Correspondence and requests for materials should be addressed to C.-H.K.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.