IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 27 October 2022, accepted 7 November 2022, date of publication 10 November 2022,
date of current version 18 November 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3221457

RESEARCH ARTICLE

Recognizing Edge-Based Diseases of Vocal Cords by Using Convolutional Neural Networks

CHEN-KUN TSUNG[1], (Member, IEEE), AND YUNG-AN TSOU[2,3,4]
[1]Department of Computer Science and Information Engineering, National Chin-Yi University of Technology, Taichung 411030, Taiwan
[2]Department of Otolaryngology-Head and Neck Surgery, China Medical University Hospital, Taichung 404333, Taiwan
[3]School of Medicine, China Medical University, Taichung 404333, Taiwan
[4]Department of Audiology and Speech-Language Pathology, Asia University, Taichung 413305, Taiwan

Corresponding author: Yung-An Tsou (d22052121@gmail.com)

This work was supported by the Ministry of Science and Technology, Taiwan, under Grant MOST 109-2221-E-167-030-MY3 and MOST 111-2314-B-039-058-MY3, and DMR-112-041 from China Medical University, Taiwan.

ABSTRACT During clinical consultations and case training, doctors analyze numerous images and sounds. A high-pressure consultation environment can increase the probability of a doctor making incorrect inferences regarding vocal cord (VC) disease. Therefore, this study applied deep learning to design an edge-based VC disease detection system (EVC-DD) for common VC conditions (e.g., nodules, polyps, and cancer) to assist doctors in conducting consultations and case studies and in verifying the consistency of their disease inferences. Through deep learning, the model extracted and recorded clinically confirmed information in its disease inference model. The experiment data set comprised videos of nodules, polyps, and cancer that were used to evaluate the performance of the proposed model. From 13 cases confirmed by two doctors, 1740 images were extracted from 13 case videos and used in the experiment. In total, 1044 (60%), 348 (20%), and 348 (20%) images were randomly obtained through five-fold cross-validation for training, validation, and testing, respectively. During the model training process, the EVC-DD model achieved 100% accuracy in detecting the three conditions required for optimal experiment results. For the results in the analysis of the independent test data with optimized configuration. the EVC-DD model achieved 99.42%, 99.42%, 99.42%, 99.42%, 98.91%, and 0.9957 for averaged F1 score, averaged recall rate, averaged precision, accuracy, Matthews correlation coefficient, and area under the curve, respectively. The EVC-DD model required only 400 s to complete its training using 1740 images. The results indicate that the inferences of the EVC-DD model were highly consistent with the results of the clinical examination by doctors and that its training was data- and time-efficient, thereby allowing the model to learn new cases quickly. Thus, the EVC-DD model can assist doctors in consultations and case analyses by providing reliable disease inferences and real-time input regarding new case knowledge.

INDEX TERMS Convolutional neural networks, vocal cords, disease recognition, nodules, polyp, cancer.

I. INTRODUCTION

Vocal cords (VC) are essential tissues that enable humans to make sounds, and changes to the VC tissue of an individual directly affect their voice. For example, nodules [25], [26], [27] and polyps [28], [29], [30] are common VC diseases that affect an individual's voice. Although nodules and polyps share similar features, doctors may encounter difficulty detecting them during a clinical examination, especially when these diseases are still in their early stages.

During a clinical examination, a doctor uses a stroboscope to access the glottis and examine VC tissue. Stroboscopes are widely used in various domains to observe real tissue images [8]. Figs. 1 and 2 display captured stroboscopy images of nodules and polyps, respectively. The areas marked by white rectangles are key features that can be used in disease inference. The small, hard, and white collagen fiber depositions on the VCs are the key features of nodules, which are masses that form mostly because of VC overuse. Polyps are

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

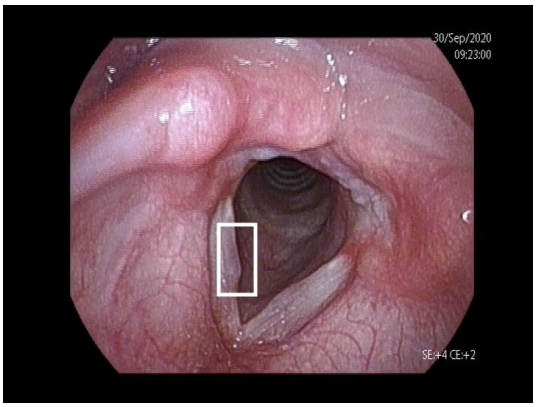


FIGURE 1. Vocal cords with nodules.

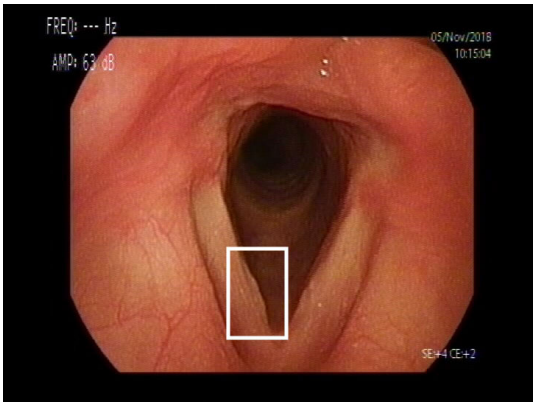


FIGURE 2. Vocal cord with polyp.

diagnosed by the presence of large, soft, red, nonfibrotic, and mucoid tissue.

Nodules and polyps cause protrusions to appear on the edges of VCs. A normal VC curve should be flat; therefore, nodules can be identified by analyzing the flatness of a VC curve. In the present study, nodules and polyps are referred to as edge-based VC diseases (EVCDs). Nodules and polyps are not distinguished by the tissue protrusions at the edge of VCs but by their degree of protrusion, color, and hardness. During a clinical examination, a doctor first determines whether the VC curve is flat; when an abnormality is detected in a patient's VC tissue, it is then classified as a nodule or polyp on the basis of its tissue characteristics. In a clinical observation, straight VC edges indicate healthy VCs, whereas nonstraight ones indicate unhealthy VCs.

Deep learning (DL) techniques provide highly accurate object and feature detection [9], [22], [23], [24], [40], [41], [51], [52], [53]. In DL models, object properties are learned from a set of labeled samples, and the presence or absence of target features in a set of input data is assessed and reported. A DL model extracts key features from target data and applies them in an inference process. Because DL is a general method for distinguishing target objects, we designed a DL-based solution for detecting EVCDs.

The present study applied DL techniques to design an EVCD detector (EVC-DD) that doctors can use as a disease inference tool during clinical examinations. This detector provides flexibility for constructing low-data models. In addition to nodules and polyps, VC cancer [9], [22], [23], [24], [40], [41] was considered as a target EVCD in the present study. The prevalence of VC cancer has increased, and the abnormal growth of VC cancer cells causes irregular shapes to form in VC curves. Therefore, VC cancer is also a type of EVCD. Fig. 3 displays the stroboscopy image of cancer. In the image, atypical hyperplasia (marked by a white rectangle) can be seen on one side of the VCs. During their early stages, VC cancer, polyps, and nodules differ slightly in terms of their symptoms; however, in stage two or three of VC cancer, hyperplasia becomes noticeably larger relative to polyps and nodules. Because VC curves account for only 0.3%–7% of an image, small-area feature recognition is the main detection feature applied in the proposed detector. The proposed EVC-DD model uses two sets of convolution layers and max pooling to obtain small-area features, thereby achieving a high recognition rate with respect to the observation of small-sized features. Because focusing on small-sized features reduces accuracy, cases with noise must be included in the model training phase to account for other nonessential features (e.g., mucoid, hyperemia, and swelling can cause EVC-DD noise).

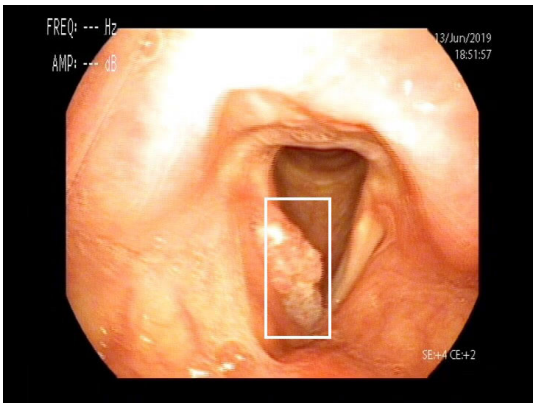


FIGURE 3. Vocal cords with cancer.

The data used in the present study were reviewed by the Research Ethics Committee of China Medical University & Hospital, which approved the use of 15 cases. The stroboscopy videos recorded by doctors during patient examinations were used, and more than 30,000 images were extracted from 13 selected stroboscopy videos. Subsequently, 1740 images were extracted from the 13 EVCD cases and included in the performance evaluation process of the present study. All stroboscopy videos and images were confirmed by two otorhinolaryngologists to ensure that the selected images represented real positive cases. In our experiment, we evaluated three convolution depths to determine how convolution depth affects the accuracy and flexibility of the proposed EVC-DD when new cases were added for training. To ensure that all images were evaluated, five-fold cross-validation was applied in the present experiments. The experimental results indicate that the EVC-DD model performed favorably for categorical accuracy; when an optimized hyperparameter

configuration was applied, the model achieved 100% accuracy. The experiment results obtained from the test data reveal that the proposed EVC-DD model achieved a similar performance during five-fold cross-validation testing, especially for cancer and in terms of F1 score, recall rate, precision, accuracy, Matthews correlation coefficient (MCC), and area under the curve (AUC). To compare the performance of the EVC-DD model against the performance of other widely applied methods, the performance of VGG16, EfficientNet, and Inception V3 models were also assessed in the present study.

The experimental results show that the performance gap between the VGG16 and EVC-DD models was <1% for nodules and cancer, and the EVC-DD model outperformed the VGG16 model by 2.63% for polyps. However, the structure of the VGG16 model is deep and complex, and that of the proposed EVC-DD model is simple and lightweight. The performance of the VGG16 model approached that of the proposed EVC-DD, but its training cost was greater than that required for the EVC-DD model. Thus, the EVC-DD model was more efficient than the VGG16 model for new EVCD cases. By contrast, the EfficientNet and InceptionV3 models were efficient training models but underperformed against the EVC-DD model for EVCD detection. Additionally, the EVC-DD model requires only approximately 400 s to train, which is a sufficiently short training time for new cases. With its excellent categorical accuracy, the EVC-DD model can help doctors to perform diagnoses and verify the diagnoses of other doctors, thereby reducing the risk of misdiagnosis. In addition, the short training time of the model enables doctors to input new characteristics related to EVCDs or cases in real time during training or case discussions to improve the model's feature comprehension.

The rest of the article is organized as follows. Section II displays the related results in object and disease detection, and a summary table is given for comparing common approaches. Section III provides the target scenario and the proposed EVC-DD model. The performance evaluation, hyperparameter adjustment, and the performance comparison between the proposed EVC-DD model and other state-of-the-art approaches are discussed in Section IV. Eventually, the conclusion and future works are stated in Section V.

II. RELATED WORKS

Artificial Intelligence (AI) methods have contributed considerably to recognition techniques and helped to improve VC disease recognition. Common AI methods include machine learning and DL. Machine learning methods classify data by attribute (e.g., gender, age, and patient's behavior), whereas DL methods classify images on the basis of their features. Davaris et al. combined a polynomial kernel and a k-nearest neighbor algorithm with a support vector machine to assess vocal fold leukoplakia [42]. Zhao et al. proposed a hybrid method involving the application of a convolutional neural network and transfer learning for the classification of VC lesions [43]. Low et al. compared the performance of several machine learning methods (e.g., logistic regression, random forest, and the Stochastic Gradient Descent classifier) in detecting unilateral vocal fold paralysis [44]. To trace the sound generation and evaluate the health of VCs, Yousef et al. applied an unsupervised machine-learning method and an active contour modeling technique to identify the position of the glottis to understand VC actions [45]. Salem et al. applied various DL networks with an Adam optimizer to detect ocular diseases [51]. Luo et al. used DL networks and proposed a novel mixture loss function for detecting various eye diseases [52]. Rath et al. adopted a long short-term memory, generative adversarial network to increase the accuracy of heart disease detection [53].

The accuracy of DL-based image recognition has considerably increased through the implementation of pixel-level feature comparison techniques. DL is used to examine the pixel structure of a given area, and the accuracy and efficiency of feature learning is increased through the operations of multiple layers. For example, autopilot systems have a high commercial development value, and in autopilot research, topics such as road detection, pedestrian detection, and obstacle avoidance have been extensively studied. Because roads contain complex information, the filtering of irrelevant information to extract crucial information is a highly relevant topic. Convolutional neural networks (CNNs) can be used to develop image recognition models for making inferences pertaining to a single disease and for classifying multiple diseases [9], [10], [11].

Fully convolutional networks (FCNs) can detect normal objects (e.g., pedestrians and moving vehicles) and semantic objects (e.g., stationary objects) [11], [12], [13]. Although the excellent segmentation ability of FCNs facilitates object detection, network correction is required to increase the applicability of such networks. Various step sizes can be integrated to achieve detailed segmentation, and global and local FCNs can be implemented to adapt to the characteristics of local and global information. FCNs are suitable for images that contain a large amount of information.

The semantics of VC images are simple. The structure of the glottis is fixed; unless cancer or a tissue mutation is present, the images obtained through a given method usually yield highly similar results, particularly during the early stages of a disease. The semantic information of VC images is less complex than that of autopilot systems; however, the subtle changes in VC images may indicate highly valuable and crucial features. Therefore, networks that rely on semantic analysis require the implementation of highly detailed observations [13], [14], [15]. The U-structure of a U-Net allows for connections to be skipped such that this requirement for medical images can be met and a high level of accuracy can be achieved with a small amount of sample data for applications such as cancer cell detection. A U-Net can perform object detection and segmentation; however, the detection of VC diseases in the glottis does not require the detection of VCs because a doctor only has to confirm that a VC curve exhibits the features of a target disease. Therefore, when a U-Net is

TABLE 1. The summary of current common DL approaches.

DL Approaches	Advantages	Applications	Issues in EVCDs
FCN	High resolution in the images that contain a large amount of information	Semantic object detecting	The improvement of accuracy in the images with simple semantics, e.g. the detection of glottis images
Unet	High accuracy with small amounts of sample data	The tissue detection for small data	The reduction of the noise generated by various observation angles, e. g. the strobiscopy images
CapsNet	Checking the position relationship between target features	Detecting the images with position-based features	The improvement of accuracy in the images with simple semantics, e.g. the detection of glottis images
R-CNN	Sharing computation with convolutional features \ through region proposal networks	Small objection detection	The reduction of the flexible during repeated training, e.g. the training time of building the model
YOLO	High inference efficiency	On-line object detection	The increase of considering the images with new features, e. g. improving the model by the stroboscopy images with new features
LeNet-5	High inference efficiency	Detecting the tissue variation	The improvement of inference performance, e.g. the accuracy
VGG16	High accuracy	Detecting general objects	The reduction of training efficiency, e.g. the training time

applied, the observation results obtained from multiple angles can generate noise that reduces accuracy. Thus, algorithms that can recognize small-area features are crucial for detecting EVCDs.

The position relationship between features is another factor for DL. For example, images with corrected features may be placed in appropriate positions, and a DL model may check all target features but fail to examine position relationships [32], [35]. CapsNet, which considers PrimaryCaps to check the position relationships between target features, is useful for recognizing images with similar features. The accuracy and efficiency of remote sensing image scene classification [33], text classification [34], and emotion recognition [36] can be improved through CapsNet. Because the collected glottis images were simple and the collected stroboscopy images were carefully collected by otorhinolaryngologists, evaluating the position relationships between tissues is unnecessary.

The abnormal tissue area is small for specific diseases; therefore, distinguishing small objects with specific features is crucial. No strict definition is applied for small objects. A small object is usually defined as an object with a resolution of 32 pixel by 32 pixel or an object that is regarded as ''small'' relative to the size of the target object in an image. Because of their high convolution cost, region-based convolutional neural networks (R-CNNs) are highly inflexible during repeated training [16], [17], [18]. Fast R-CNNs share computational load with convolutional features through region-based networks; thus, their computation efficiency and accuracy can be increased.

Major changes were made to the fourth version of the YOLO model [19], which uses CSPDarknet53 + PANet + SPP to reduce computation time and combines local and global features to improve the expression ability of feature maps. Different methods use different entry points to identify small objects, but they are generally highly applicable in practice. However, the flexibility of model training is crucial in medical diagnosis. Doctors should add cases to a model when they encounter indicative VC tissue to ensure that the experience associated with a case is captured by the model. Therefore, the flexibility of repeated training is a key consideration for network designers.

Borsato et al. marked the edge of eyeballs by object color and shape to enable the evaluation of user focus through the tracking of eyeball positions [8]. Eyeball tracking is tractable because eyeball images can be stored directly and an invasive test is not required. However, VC disease inference is more difficult to implement relative to eyeball position tracking because patients may feel uncomfortable. Vo et al. proposed a hybrid model that uses CNN, random forests, and support vector machines to identify cancers [9].

Several major frameworks may be applied to detect EVCDs. LeNet-5 proposed by LeCun is a back-propagation network [37], and easy use is the major advantage of LeNet-5 that the image preprocess is not too much. LeNet-5 can achieve high recognition rates, especially for the recognition of handwritten and machine-printed characters. LeNet-5 has a seven-layer structure with three convolutional layers, two pooling layers, one fully connected layer, and one output layer. Some layers do not fully use the information from preceding channels to increase computational efficiency. However, the appearance of target features in unused channels reduces the recognition rate of a model. Vo et al. applied an LeNet-5 model and used images captured through confocal laser endomicroscopy to distinguish healthy tissues from cancerous VC tissues [9]. Analyzing the properties of tissues is another method for implementing disease recognition, and tissue variations can be detected in the early stage of a disease. However, the analyzer should have sufficient knowledge about the relevant tissue properties. For real-time inferences, the method proposed by Vo et al. yields a recognition rate of >81.5%. Vo et al. also reported that increasing the search area of a model increases its accuracy as well as its computational cost. Cho and Choi compared the performance of the CNN6, VGG16, Inception V3, and Xception models in terms of recognizing nodules and vocal fold granuloma [38]. The simulation results reported by Cho and Choi indicate that the VGG16 model has a 99% recognition rate, which is

high and can be attributed to the large-scale network structure of the model. Because the VGG16 model uses 138,357,544 parameters, it requires a lengthy training time [39]. When valuable cases are used to train the VGG16 model, its network structure should be modified to reduce training time; further studies on recognition rate are required.

The aforementioned DL methods for detecting EVCDs and the related studies are listed Table 1, which summarizes their advantages, applications, and challenges in the context of EVCD detection. Because glottis images are simple and the objects found in such images are not complex, the detection of semantic objects is unnecessary; however, the training efficiency associated with new features and accuracy should be considered.

III. SCENARIO AND METHODS

A. OBJECTIVE OF THIS STUDY

The present study focused on EVCDs (i.e., nodules, polyps, and cancer); specifically, the case managers of this study collected data pertaining to several cases for each target disease. All the cases were confirmed by two otorhinolaryngologists, and those with obvious key features were selected. The data used in the present study were reviewed by the Research Ethics Committee of China Medical University & Hospital, and 15 cases were approved. Subsequently, 13 cases were evaluated in the present study; Table 2 presents the cases selected for evaluation.

The objective of the present study was to design a detection model that can recognize EVCDs in VC tissues by examining the stroboscopy videos recorded during clinical examinations. Otorhinolaryngologists can use the proposed AI model to verify the appropriateness of their disease inferences. Moreover, doctors can input new cases into the model, thereby enabling the model to consider newly discovered key features of VC tissues and minimize the cost of model training.

B. SYSTEM DESIGN

The entire process of designing and implementing the EVC-DD is illustrated in Fig. 4. Here the steps are briefly described as follows.

1) The otorhinolaryngologist used Stroboscope to observe and record patient's VC status during the consultation, and the diagnosis videos are collected.
2) The engineer used Free Video to JPG Converter[1] to convert videos to images at 50 Hz.
3) All images are carefully examined by two otorhinolaryngologists to make sure that each image has correct features of target diseases.
4) All examined images are randomly classified into training data and testing data for model training and performance evaluation.
5) The training data were used to train the proposed deep learning model.

[1]https://free-video-to-jpg-converter.en.uptodown.com/windows

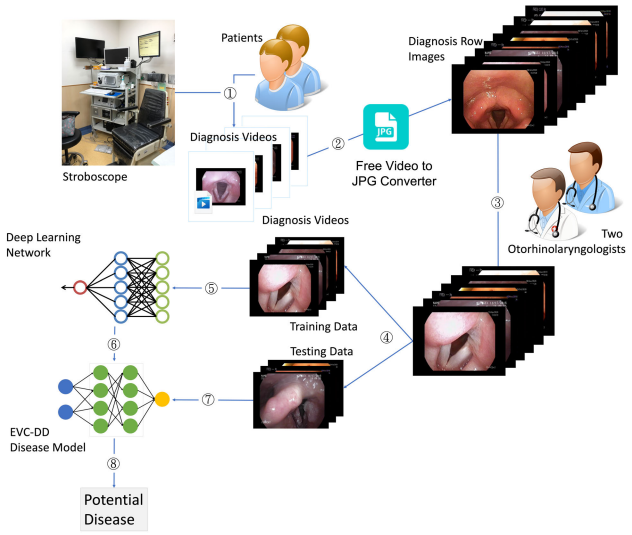


FIGURE 4. The overall processes of design and implementation of EVC-DD.

TABLE 2. Distribution of cases by condition.

Disease	Evaluated cases	Selected cases
Nodules	7 (36.84%)	5 (38.46%)
Polyp	3 (15.79%)	3 (23.08%)
Cancer	9 (47.37%)	5 (38.46%)

6) The EVC-DD disease model could be derived after training model.
7) The testing images were used to evaluate the performance of the derived model.
8) The derived model outputted the potential disease for each given testing image.

After step six, the detection model is obtained and can then be used by doctors. The disease inference process is as follows:

1) Uploading VC video: the VC video is uploaded, and the VC images are captured for examination by the EVC-DD model.
2) Disease inference: the EVC-DD model infers the diseases in each image one by one.
3) Report: the results of disease inference are summarized, and an analysis report of the disease is generated.

C. NETWORK DESIGN

After the stroboscopy images and VC features were confirmed by the two otorhinolaryngologists, we identified the following network design issues:

1) Plane image analysis: EVCDs can be detected by identifying the deformation of VC edges in an image. Doctors identify VC diseases during clinical examinations by checking for deformations on VC edges. Therefore, a two-dimensional (2D) image analysis is appropriate.
2) Slight changes in VC edges: Polyps only cause small changes in VCs. A sampling analysis revealed that

polyp pixels account for only 0.3%–0.7% of an image area. Therefore, a small-area image analysis must be performed.

3) Enhancement of object features and reduction of data redundancy: Because nodules cause more swelling relative to polyps and an entire VC edge may swell when nodules form, nodule features account for approximately 2.5%–7% of an image area. Therefore, enhancing the key points of features is an effective method for detecting nodules.
4) Classification of three conditions: Because three conditions were evaluated in the present study, multiclass classification was considered for achieving machine vision (MV).

On the basis of the aforementioned requirements, a network with the following elements was designed:

1) A 2D convolution layer with a small kernel was implemented to convolve the features in each image and identify small-area changes. Because stroboscopy images are generally similar, notable differences can only be observed when cancer causes symptoms to appear on the tissue near the glottis. Therefore, small kernels are useful for analyzing VC images.
2) Max pooling is suitable for large feature differences, and it can increase judgement accuracy during the evaluation of a series of cases. Therefore, max pooling can be used to achieve high-accuracy nodule detection.
3) A dense layer can be used for multiclass situations. Before dense layer classification is performed, data are reduced to a single dimension through layer flattening.

Fig. 5 illustrates the network structure of the EVC-DD model. Continuous and precise feature recognition was achieved by implementing two 2D convolution layers and max pooling. The disease class of the image was determined by a flattened layer and a dense layer. This network structure was named "VC-MV_{21}" because of its two 2D convolution layers and one-layer convolution depth. "VC-MV_{21}" matches the VC edge features of polyps and early-stage nodules. However, its recognition efficiency may be less than ideal for large-area, mid-to-late-stage nodules or cancer. Therefore, we applied two- and three-layer depths for each set of 2D convolution layers to identify the continuous feature of large areas; the models with two- and three-layer depths were named "VC-MV_{22}" and "VC-MV_{23}", respectively. Figs. 5(a) - 5(c) illustrate the network structures of "VC-MV_{21}", "VC-MV_{22}", and "VC-MV_{23}" network structures, respectively. Convolution layers were added to each of the two 2D convolutions to increase the depth of recognition, thereby increasing the dimension of features and the number of pixels considered in large-area features.

D. HYPERPARAMETERS SETTING

Table 3 presents the hyperparameter settings. The principle behind the complex hyperparameter settings is explained below. The learning environment of the EVC-DD model was allotted 16 GB of memory space. Therefore, the batch size can be increased to 32 to simultaneously interpret more information. Setting epoch at 20 is sufficient for EVC-DD learning, and epoch can be reduced to approximately 16 in early stop. Because EVC-DD uses smaller kernels, the kernel size can be set to 3 by 3. Activation uses reLU in each convolution layer, whereas the dense layer uses softmax; because the characteristics of the features are emphasized in the learning process, reLU can yield high-quality results with few resources. The final determination of condition in the image based on the probability distribution of each condition. Therefore, softmax was suitable for the multiclass classification problem. The one-hot format was considered for labeling. Because the resolution of Stroboscopy images is 540, 720, input_shape was set to 540, 720, 3. Categorical crossentropy was applied to the EVC-DD model because three conditions were considered. Metrics were set to categorical accuracy.

TABLE 3. Hyperparameters setting.

Hyperparameter	Configuration
batch size	21
Epoch	20
kernel size	3 by 3
Activation	reLu in each convolution Layer while softmax in Dense Layer
Labeling	one-hot format
input_shape	540, 720, 3
Loss function	categorical cross entropy
Metrics	categorical accuracy

E. ENVIRONMENT

The EVC-DD platform specifications for training, validating, and testing the model is as follows: Intel Core i9, 64 GB Memory, 1TB SSD, 1Gb Intel Ethernet, NVIDIA GeForce GTX 1060.

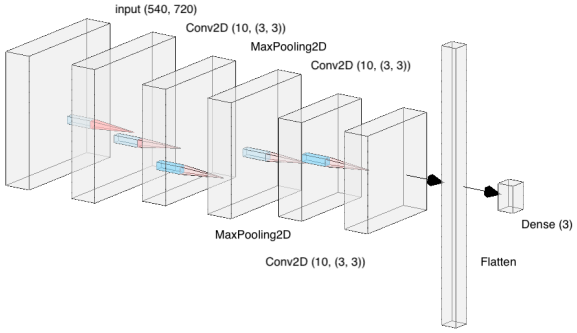
F. IMAGE DISTRIBUTION

We collected a total of 1740 images to identify polyps, nodules, and cancer. The distribution for each condition in the different tasks was 60% (1044 images) for training, 20% (348 images) for validation, and 20% (348 images) for testing, and one-hot encoding was used to label the conditions.

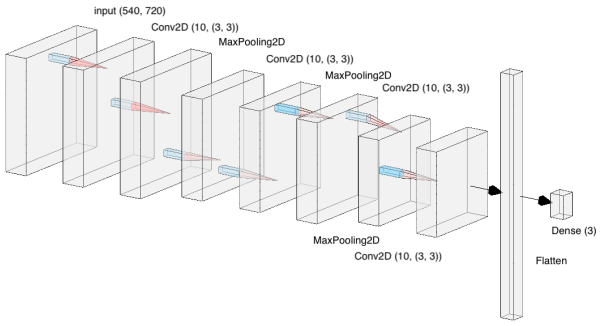
IV. PERFORMANCE EVALUATION

A. DATA SET

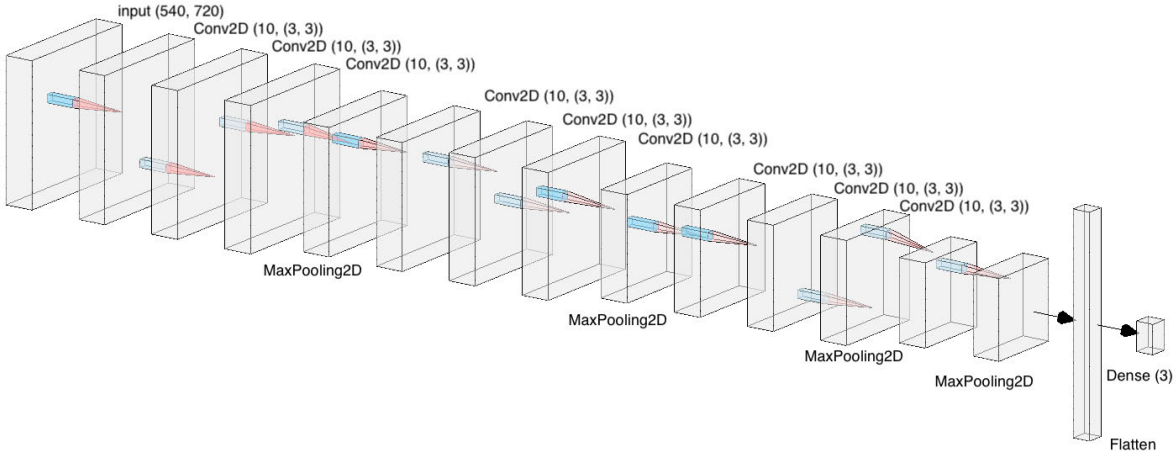
The data set used in this study comprised videos recorded by otorhinolaryngologists from China Medical University Hospital; these videos were recorded during regular examinations of the VC tissues of patients. During a regular examination, a doctor uses a stroboscope to examine the VC tissue of a patient and record the full examination process. In accordance with the processes described in Section III-B, the two otorhinolaryngologists confirmed that the 19 videos indicated the presence of the target EVCDs. Next, we extracted the frame-by-frame images from all videos and examined them to verify that the VC tissues in these images exhibit the key features of



(a) Network structure of VC-MV$_{21}$.



(b) Network structure of VC-MV$_{22}$.



(c) Network structure of VC-MV$_{23}$.

FIGURE 5. The proposed network structures.

the target EVCDs. All the selected images were then labeled on the basis of the target EVCD that they indicated and used to evaluate the performance of the proposed network.

When a doctor uses a stroboscope to observe VC tissue, three steps are performed; specifically, the stroboscope (1) enters the nasal cavity, (2) captures the images of VC tissue, and (3) then leaves the nasal cavity. Because the first and third steps do not involve the examination of VC tissue, only the data captured during the second step were used in the present study.

After the stroboscopy videos were obtained, we used image capture software to obtain images from the videos at 50 Hz. Each image was inspected to ensure that they pertained to the second step of the examination process. To capture the movement of throat tissue, the sampling rate of the stroboscope is usually high. Therefore, capturing images at 50 Hz provides a balance between obtaining a clear view of tissue changes and maximizing the number of images obtained.

The glottis can either be open or closed (Figs. 6 and 7). When it is closed, the shape of VC edges becomes unclear, and identification of the features of nodules and polyps becomes difficult. Therefore, we only used images that show an open glottis. Stroboscopy can be performed in both flash and nonflash modes (Figs. 8 and 9), enabling doctors to observe the movement of VCs under different lighting conditions. Because doctors usually confirm the state of VCs in both modes, both flash- and nonflash-mode images were used in the present study.

Fig. 4 presents a flowchart describing the collection of the 19 stroboscopy videos and the candidate cases that they corresponded to. A total of 13 cases were selected from 19 candidate cases because of video quality considerations, and more than 30,000 images were extracted from the 13 videos that corresponded to the 13 cases. Eventually, 1740 images were selected after the collected images were evaluated frame by frame. Table 4 lists the number and proportion of the images that correspond to each condition. The selected images all displayed an opened glottis and clearly visible target tissues and VCs.

B. HYPERPARAMETER EVALUATION

When more layers are used in a network structure, the number of parameters increases, which also increases the model training time. However, the disease inference time is not notably affected. This study designed the model to satisfy clinical consultation needs and doctors' case training needs to ensure the model training time would not affect the clinical consultation or the quality of case training. Therefore, the

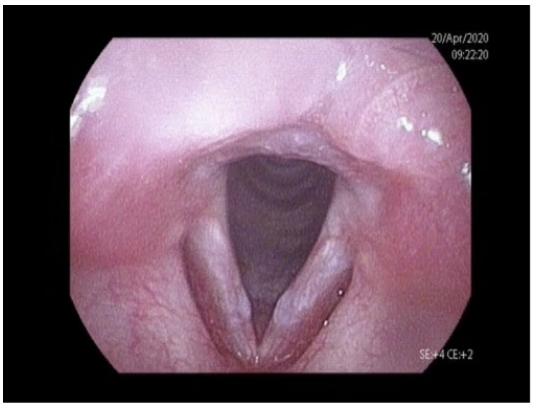


FIGURE 6. Open glottis.

FIGURE 7. Closed glottis.

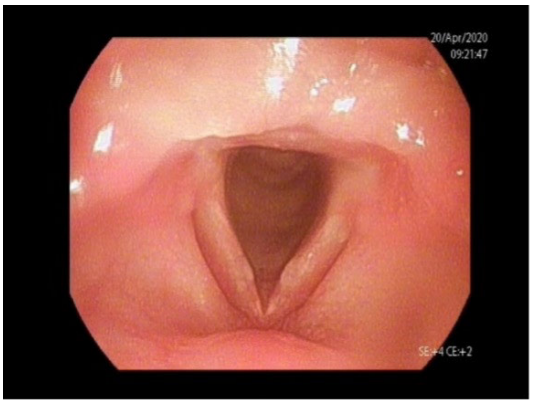


FIGURE 8. Non-flash mode.

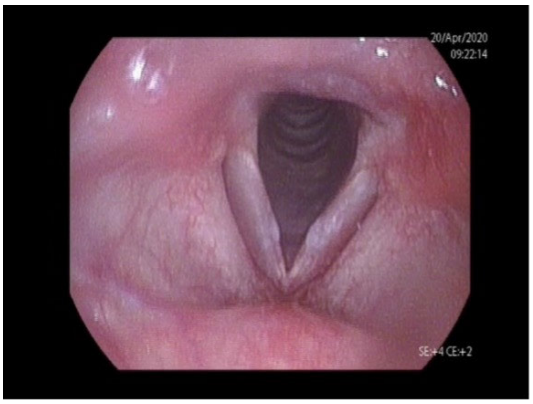


FIGURE 9. Flash mode.

accuracy of the disease inference was prioritized over the time cost of model training. The performance of the different networks was evaluated in terms of accuracy and loss.

1) Accuracy: because this study investigated three EVCDs, multiclass judgement accuracy would directly benefit doctors the most. Therefore, the model was evaluated in terms of categorical accuracy. The prediction error and results can be obtained through categorical accuracy.
2) Loss: the difference between the identification of the conditions during training, validation, and testing and the actual occurrence of the condition. The closer the loss to 0, the higher the quality of the disease inference.

Figs. 10 and 11 display the categorical accuracy and loss of each network in the identification of the conditions, respectively. The x-axis is the learning rate, and the y-axis are the categorical accuracy and loss. The three models achieved the highest categorical accuracy and the lowest loss. When the learning rate was 0.01, the loss and categorical accuracy of models 1 and 2 were approximately 0.001 and 0.99, respectively. VC-MV_{23} performed better when its learning rate was 0.001; its corresponding categorical accuracy and lowest loss were approximately 0.07 and 0.98, respectively. In terms of the overall performance, all three networks achieved near-optimal categorical accuracy and loss, but VC-MV_{23}, which had the largest network depth, required a lower learning rate to achieve results similar to those of models 1 and 2. Therefore, models 1 and 2 outperformed VC-MV_{23} in disease inference.

These results were yielded when epoch was 20. This study evaluated disease inference performance under different epoch settings. Figs. 12 and 13 present the performance of the three models under a fixed learning rate of 0.00001 and epoch at 20, 50, and 100. The performance of the models improved with the increase in epoch; with epoch at 100, the difference in categorical accuracy among the models was less than 0.05, and the loss was only 0.05. Although the difference in performance was unsubstantial, the performance of the network with the largest depth was inferior to that of the shallow networks. The effect of network depth on disease inference was not as considerable as expected, but the models trained by the shallow networks performed better in EVC-DD.

We further analyzed the differences in performance among the three models. Figs. 14 and 15 present the results for categorical accuracy and loss, respectively, with extreme values of learning rates of 0.0006 and 0.0001 and epochs of 20 and 50. The x-axis is a combination of different hyperparameters;

TABLE 4. Number and proportion of images by condition.

Disease	Image Count	Distribution
Cancer	486	27.93%
Nodules	1072	61.61%
Polyp	182	10.46%

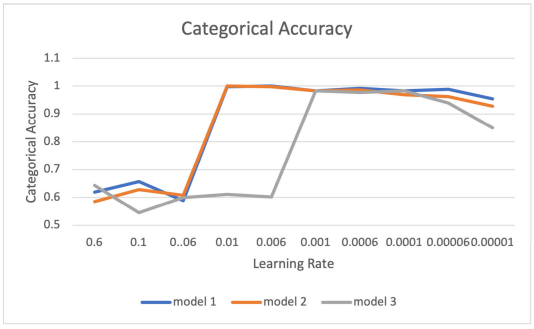


FIGURE 10. Disease Inference results in categorical accuracy in various learning rate settings.

FIGURE 11. Disease inference results in loss in various learning rate settings.

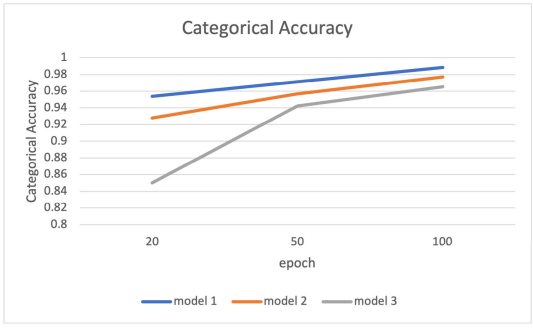


FIGURE 12. Disease inference results in categorical accuracy in various epoch settings.

the first number is the learning rate, and the second number is epoch. The results indicated that the performance with a learning rate 0.0006 was superior to that with a learning rate of 0.0001 and that performance was superior with an epoch of 50. VC-MV_{23} performed well when the learning rate was 0.0001 and epoch was 50, but its performance was inferior to that of the other models.

In the following experiments, we consider 5-Fold Cross-Validation [50] to guarantee that all images are evaluated. The validation strategy applied in this experiment is shown in Fig. 16, and we derived five experiment results. In the beginning, all images are shuffled randomly, and the shuffled images are classified into five groups with a sequence. Each group will be training dataset thrice, validation dataset once, and testing dataset once to make sure that each image will be evaluated.

C. DISCUSSION

1) TRAINING TIME

High categorical accuracy and low loss were achieved when the hyperparameters were more detailed for long-term training. However, the benefit of increasing the accuracy and decreasing the loss for extreme configurations with a low learning rate and large epoch was marginal (Figs. 14 and 15). Because new cases added to model training can increase the inference accuracy of the model, the hyperparameter settings for periodic model retraining are key considerations in practice. The analysis of training time revealed that the time required to train the model was proportional to the epoch but was unrelated to the learning rate. Table 5 presents the time required to train the model under different epoch values; the training time of each model can be determined on the basis of epoch. A learning rate of 0.0006 and an epoch of 50 can achieve a balance between performance and practicality. Additionally, the difference in performance between models 1 and 2 was only 0.586% with these hyperparameters (VC-MV_{21}: 0.9885; VC-MV_{22}: 0.9827). Therefore, VC-MV_{21} was superior in terms of training time.

The training results of VC-MV_{21} with learning rate 0.006 and epoch 50 are shown in Figs. 17 and 18 for loss and categorical accuracy. The EVC-DD model has plenty of epochs for convergency, so the tails are stable for both training and validation curves. Moreover, in the first 10 epochs, the curves drop dramatically and smoothly. Model 1 is the shallow network structure, so the convergency curves do not require too many epochs to reach steady states, and this property helps in the training time of modifying the EVC-DD model in considering new cases.

2) PREDICTIONS FOR EACH CATEGORY

As highlighted in Section III-F, 20% of the images were selected randomly for use as the test data for each target EVCD. Therefore, the EVC-DD model examined 36, 214, and 97 images showing polyps, nodules, and cancer, respectively (Table 4). To measure the model's performance for all images and verify that all images were evaluated, we performed five-fold cross-validation [50]. Through the validation strategy applied in this experiment (Fig. 16), we obtained five sets of experiment results. The confusion matrixes of three iterations are presented in Fig. 19. Our predictions for cancer classes were accurate, but those pertaining to several polyp and nodule images were incorrect. From specific

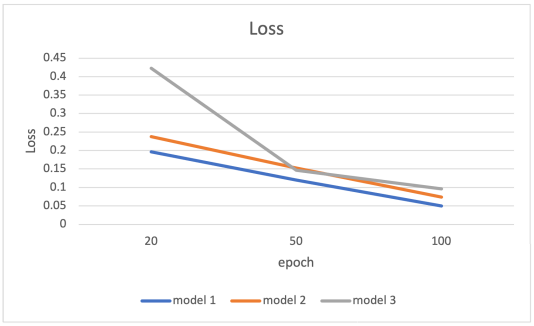


FIGURE 13. Disease inference results in loss in various epoch settings.

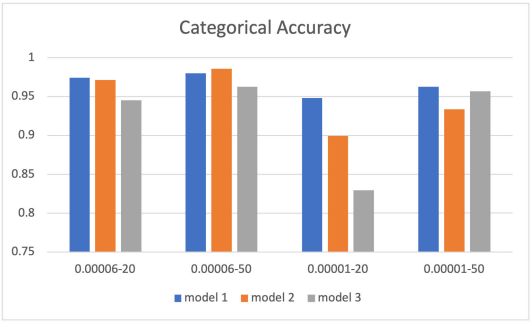


FIGURE 14. Disease inference results in categorical accuracy in various epoch settings.

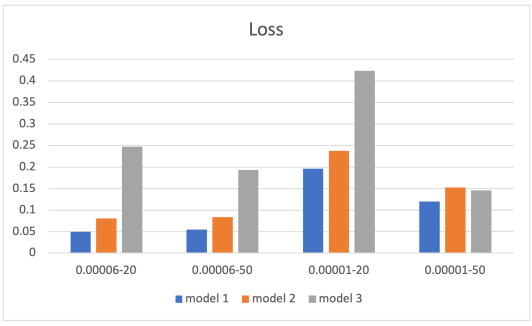


FIGURE 15. Disease inference results in loss in various epoch settings.

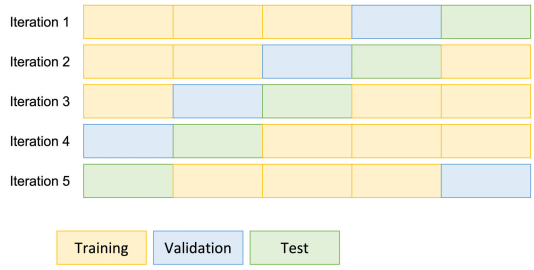


FIGURE 16. The proposed 5-Fold cross-validation.

stroboscope angles, cancer features appear different from those of polyps and nodules; by contrast, the difference between the features of polyps and nodules is small. Clinical experience indicates that the tissue properties of polyps are similar to those of nodules; thus, predictions pertaining to polyps and nodules are occasionally inaccurate. Moreover, the number of incorrect predictions for polyp and nodules matched the image distribution; thus, the results obtained were reasonable (Fig. 19).

On the basis of the results presented in Fig. 19, we used Python and Sklearn to calculate the performance of the proposed EVC-DD model (see Table 6 for the results), which was revealed to be stable in terms of F1-score, recall rate, and precision. The weighted performance model applied in the present study considered the ratio of the number of disease images to total images; thus, the distribution of image number was considered in the weighted performance of the model. Although the performance of the model in terms of the weighted MCC was slightly poorer relative to its performance in terms of other metrics, the model still achieved 97.83% weighted MCC in iteration 2. Clinical experience indicates that polyps and nodules are difficult to distinguish, particularly during their early stages. The EVC-DD model made erroneous inferences for four out of 250 images (error rate of approximately 1.6%); therefore, its performance in terms of the MCC was slightly poorer relative to its performance in terms of other metrics.

The AUC results of the present study are listed in the final row of Table 6, and they reveal the absence of a gap between the actual diagnoses and the predictions (the maximum value of AUC is 1). The EVC-DD model achieved high AUC values of >0.98, which were similar to the optimal results. After the results for all iterations were verified, the AUC scores were revealed to match the prediction results; iteration 3 produced the optimal result, and the final result was obtained in iteration 5. However, the miniscule gap between the optimal and final AUC scores (0.0087) indicates that the variance of the EVC-DD model's performance from iteration to iteration was low.

3) SCALABILITY PERTAINING TO COVERAGE OF OTHER VC DISEASES

The proposed EVC-DD model achieved a high recognition rate for EVCDs, which can primarily be attributed to its ability to perform DL-based feature evaluations. For images with specific and clear features, DL allows for the appropriate selection of valuable features, resulting in a high recognition rate after a prediction model is obtained. However, the scalability of the EVC-DD model for recognizing untrained diseases is low. The inference process of the EVC-DD model was dependent on the features of the training images used to train it; thus, its accuracy for predicting the images of untrained diseases is low. To overcome this problem, the EVC-DD model could be implemented with a shallow network to reduce its training time. However, a shallow network structure can only reduce the training time of the model to increase its training efficiency when new features are considered; this type of network does not improve the scalability of

TABLE 5. The running time required in the training phase for each model (unit: second).

	VC-MV$_{21}$	VC-MV$_{22}$	VC-MV$_{23}$
Epoch = 20	166	268	390
Epoch = 50	415	670	975
Epoch = 100	830	1340	1950

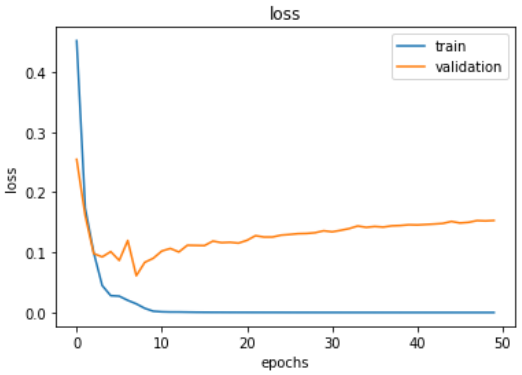


FIGURE 17. The training history of EVC-DD in loss.

the model. Because a trained EVC-DD model can achieve a high recognition rate, connecting multiple models with various trained models may increase its level of disease coverage; this strategy could be considered in future EVC-DD studies.

4) PERFORMANCE OF EVC-DD MODEL RELATIVE TO OTHER NETWORKS

We compared the performance of the proposed EVC-DD with those of other DL networks. On the basis of the results of other studies [46] and [38], the VGG16 [47], EfficientNet [48], and InceptionV3 [49] models were selected for the comparison analysis. Although researchers have studied the MobileNetV2 model [46], which can be built at a low computational cost, it was excluded from our comparison analysis because its accuracy is unacceptable for clinical applications.

Each algorithm was implemented using the simulation environment and data setting applied in the preceding experiment performed in the present study. The metrics used were F1 score, recall rate, precision, accuracy, and MCC. F1 scores, recall rates, and precision values were calculated for each target EVCD, and the weighted average was obtained; accuracy and MCC were used to assess the overall performance of the models through sklearn. The performance of the EVC-DD, VGG16, EfficientNet, and InceptionV3 models is presented in Tables 7, 8, 9, and 10, respectively. Furthermore, Table 7 lists the optimal results of the EVC-DD model, which were obtained during iteration 3.

The results obtained in this experiment were matched with those reported in other studies [38], [46]. The results of the VGG16 model were similar to those of the EVC-DD model (i.e., 100% prediction accuracy for cancer and more favorable performance for nodules than for polyps).

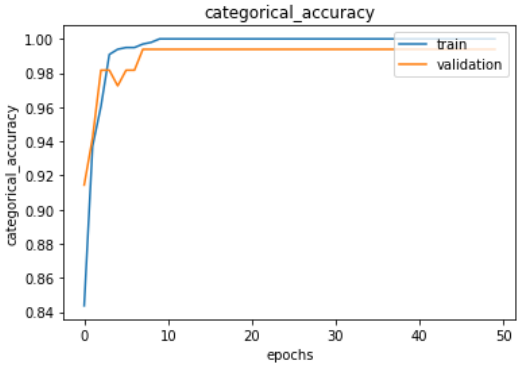


FIGURE 18. The training history of EVC-DD in categorical accuracy.

The EfficientNet and InceptionV3 models both performed most favorably for nodules, followed by cancer and polyps. The sequence pertaining to performance corresponded to the number of images of each EVCD that were used. That is, both the EfficientNet and InceptionV3 models encountered difficulties in capturing the key features of small-scale data sets. Therefore, both networks were only helpful for diseases with well-defined features. Although these two networks still achieved a performance level of >85%, they underperformed against the EVC-DD and VGG16 models.

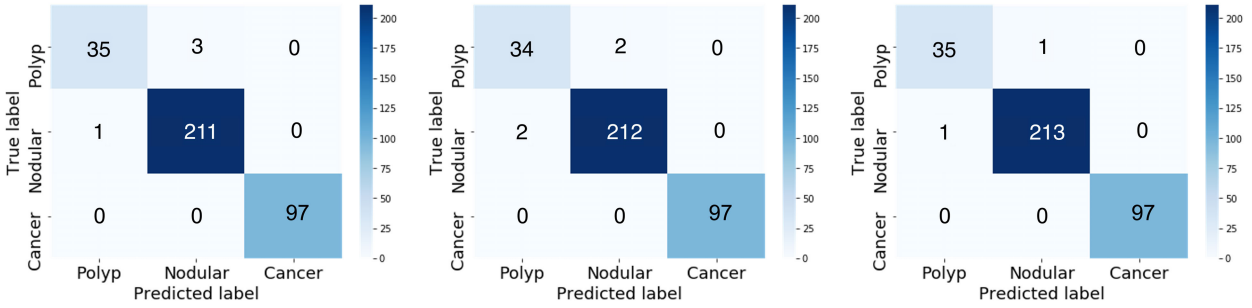
5) PERFORMANCE OF EVC-DD MODEL VS. VGG16 MODEL

The aforementioned results reveal that the proposed EVC-DD model was highly accurate in recognizing EVCDs and could be trained at a low training cost. We compared the performance of the EVC-DD model with the performance of other networks. On the basis of the results reported by Cho et al. [46] and Cho and Choi [38], we compared the performance of the EVC-DD model with the performance of the VGG16 [47], EfficientNet [48], and InceptionV3 [49] models. According to one study [38], [46], the VGG16 model can achieve a high level of accuracy (>99%); thus, we conducted further experiments to compare the performance of the EVC-DD and VGG16 models.

Cho and Choi reported that the VGG16 model is excellent at recognizing VC diseases [38]; therefore, we included the VGG16 model in our performance comparisons. To match the properties of the EVCD and VGG16 models, we performed the following modifications:

1) The original VGG16 model only accepts 224 by 224 images; thus, the size of all images was adjusted from 540 by 720 to 224 by 224.
2) The output layer of the original VGG16 model comprised 1000 dimensions, which was reduced to three because of the number of EVCDs examined in the present study.

During the simulation of parameter optimization processes, the VGG16 model requires 134,272,835 parameters but achieves a categorical accuracy of approximately 70%. The VGG16 and EVC-DD models both consider the same



(a) The confusion matrix in iteration 1. (b) The confusion matrix in iteration 2. (c) The confusion matrix in iteration 3.

(d) The confusion matrix in iteration 4. (e) The confusion matrix in iteration 5.

FIGURE 19. The confusion matrix derived by the test results of EVC-DD in each iteration.

TABLE 6. The performance of the proposed EVC-DD.

	Iteration 1				Iteration 2				Iteration 3			
	Polyp	Nodule	Cancer	Weighted average	Polyp	Nodular	Cancer	Weighted average	Polyp	Nodular	Cancer	Weighted average
F1 score	94.59%	99.06%	100%	98.83%	94.44%	99.07%	100%	98.85%	97.22%	99.53%	100%	99.42%
Recall rate	92.1%	99.53%	100%	98.85%	94.44%	99.07%	100%	98.85%	97.22%	99.53%	100%	99.42%
Precision	97.22%	98.6%	100%	98.84%	94.44%	99.07%	100%	98.85%	97.22%	99.53%	100%	99.42%
Accuracy	98.85%				98.85%				99.42%			
MCC	97.85%				97.83%				98.91%			
AUC	0.9914				0.9914				0.9957			

	Iteration 4				Iteration 5			
	Polyp	Nodule	Cancer	Weighted average	Polyp	Nodular	Cancer	Weighted average
F1 Score	94.59%	98.82%	99.49%	98.55%	91.89%	98.59%	100%	98.25%
Recall rate	92.11%	99.53%	98.98%	98.56%	89.47%	99.06%	100%	98.27%
Precision	97.22%	98.13%	100%	98.56%	94.44%	98.13%	100%	98.25%
Accuracy	98.56%				98.27%			
MCC	97.32%				96.77%			
AUC	0.9892				0.987			

TABLE 7. The best performance of the proposed EVC-DD.

	Polyp	Nodular	Cancer	Weighted average
F1 score	97.22%	99.53%	100%	99.42%
Recall rate	97.22%	99.53%	100%	99.42%
Precision	97.22%	99.53%	100%	99.42%
Accuracy	99.42%			
MCC	98.91%			
AUC	0.9957			

TABLE 8. The performance of the VGG16.

	Polyp	Nodular	Cancer	Weighted average
F1 score	95.89%	99.3%	100%	99.13%
Recall rate	94.59%	99.53%	100%	99.14%
Precision	97.22%	99.07%	100%	99.13%
Accuracy	99.14%			
MCC	98.38%			

kernel size; however, the VGG16 model has larger channels than those of the EVC-DD model. The VGG16 model considers between 64 and 512 channels, and we inferred that this model considered an excessive amount of information, which resulted in a reduction in the values of target features. Therefore, the number of channels of the VGG16 model was

TABLE 9. The performance of the EfficientNet.

	Polyp	Nodular	Cancer	Weighted average
F1 score	86.84%	94.43%	92.68%	93.01%
Recall rate	82.5%	97.99%	87.96%	93.08%
Precision	91.67%	91.12%	97.94%	87.64%
Accuracy	93.08%			
MCC	87.64%			

TABLE 10. The performance of the InceptionV3.

	Polyp	Nodular	Cancer	Weighted average
F1 score	83.54%	93.13%	88.89%	90.6%
Recall rate	76.74%	97.94%	83.64%	90.78%
Precision	91.67%	88.79%	94.85%	91.06%
Accuracy	90.78%			
MCC	83.79%			

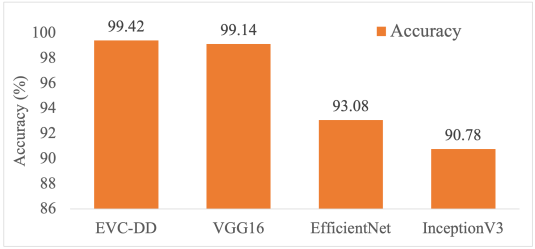


FIGURE 20. The accuracy of the proposed EVC-DD, VGG16, EfficientNet, and InceptionV3.

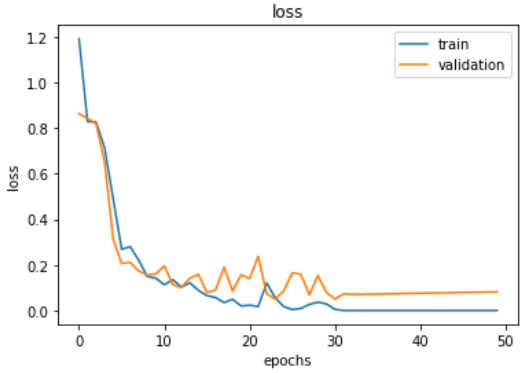


FIGURE 21. The training history of VGG16 in loss.

reduced by half for each layer, and the model's categorical accuracy was also increased from 70% to 99%. Furthermore, the number of parameters used by the VGG16 model was reduced to 16,789,411, and the model's training time was reduced to 4.5 s/epoch.

The training results of the VGG16 model are presented in Figs. 21 and 22. Both the EVC-DD and VGG16 models achieved similar recognition rates (approximately 99%); however, the VGG16 model required more epochs to reach convergency (Figs. 17, 18, 21, and 22). Moreover, the VGG16 model required approximately 4.5 s/epoch for training, whereas the EVC-DD model only required 1.5 s/epoch. Therefore, the EVC-DD and VGG16 models performed

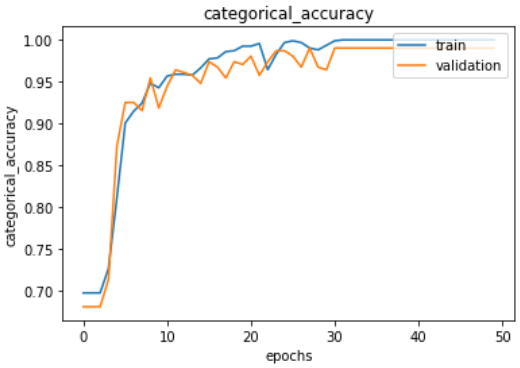


FIGURE 22. The training history of VGG16 in the categorical accuracy.

similarly in terms of recognition rate, but the EVC-DD could more quickly consider new features when new EVCD cases were input.

V. CONCLUSION

To assist doctors in observing the symptoms of patients during clinical consultations, medical equipment manufacturers have developed sampling equipment with light and shadow modes and multiple resolution and sampling-rate options. These tools greatly assist doctors in identifying disease features and formulating treatment plans. The present study applied DL to construct an EVC-DD model that records and stores lesion data and can be integrated into otorhinolaryngology practice because it does not require the collection of additional data or modification of consultation processes. The EVC-DD model can achieve an accuracy of >99%, and under specific parameter settings, a 100% accuracy can be achieved. During consultations, the EVC-DD model can help doctors to assess the consistency of their disease inferences. In case discussions, the EVC-DD model can help doctors to engage in case learning. Consequently, the experience of senior doctors can be effectively applied through the EVC-DD model to support consultations and case learning.

The EVC-DD system can achieve a high recognition rate. The integration of Stroboscope video stream could enable the model to obtain auxiliary information for disease inference in real time during clinical consultations with doctors and enable the immediate confirmation of the consistency and appropriateness of a diagnosis. Research should be conducted on the applicability of EVC-DD to conditions other than EVCDs, such as diseases related to VC oncology, VC edema, and laryngopharyngeal reflux. In the future, further research could be conducted on the convenience of EVCD and the extended sickness of other VC diseases.

The proposed approach recognizes EVCDs with high accuracy, and it will be applied to following two applications:

1) Extending the target diseases from EVCDs to other VC diseases. After obtaining some trained models for various VC diseases, we will design a mechanism to connect the trained models to recognize various VC

diseases, and the disease coverage of EVC-DD will be increased. This idea came from the research results of Yan et al. [31]. Since each model provides high accuracy for a specific disease, the hierarchical or tree testing structure would extend the disease converge. This interesting issue will be designed and implemented after finishing some ECVD models.

2) The model also enables real-time recognition during clinical consultations. The core function of the EVC-DD model, namely recognizing EVCDs, was designed and tested in the present study. In the future, the core function of the EVC-DD model will be applied to a stroboscopy system to realize real-time recognition during clinical consultations (as opposed to offline disease inference). With the support of this model, doctors can immediately receive consultation suggestions and therby reduce the risk of errors during clinical consultations.

ACKNOWLEDGMENT

The authors would like to thank the reviewers for their time and effort spent reviewing this article to improve its quality.

REFERENCES

[1] G. O. Barnett, J. J. Cimino, J. A. Hupp, and E. P. Hoffer, "DXplain: An evolving diagnostic decision-support system," *J. Amer. Med. Assoc.*, vol. 258, no. 1, pp. 67–74, 1987.

[2] T. van der Weijden, H. Post, P. L. P. Brand, H. van Veenendaal, T. Drenthen, L. A. van Mierlo, P. Stalmeier, O. C. Damman, and A. Stiggelbout, "Shared decision making, a buzz-word in The Netherlands, the pace quickens towards nationwide implementation," *Zeitschrift für Evidenz, Fortbildung und Qualität im Gesundheitswesen*, vols. 123–124, pp. 69–74, Jun. 2017.

[3] L. Perestelo-Perez, A. Rivero-Santana, J. Perez-Ramos, M. Gonzalez-Lorenzo, J. G.-S. Roman, and P. Serrano-Aguilar, "Shared decision making in spain: Current state and future perspectives," *Zeitschrift für Evidenz, Fortbildung und Qualität im Gesundheitswesen*, vol. 105, no. 4, pp. 289–295, Jan. 2011.

[4] H. H. Liao, H. W. Liang, H. C. Chen, C. I. Chang, P.-C. Wang, and C. L. Shih, "Shared decision making in Taiwan," *Zeitschrift für Evidenz, Fortbildung und Qualität im Gesundheitswesen*, vol. 123, pp. 95–98, Jun. 2017.

[5] C.-S. Wu, C.-J. Kuo, C.-H. Su, S. Wang, and H.-J. Dai, "Using text mining to extract depressive symptoms and to validate the diagnosis of major depressive disorder from electronic health records," *J. Affect. Disorders*, vol. 260, pp. 617–623, Jan. 2020.

[6] J. Lever, M. R. Jones, A. M. Danos, K. Krysiak, M. Bonakdar, J. K. Grewal, L. Culibrk, O. L. Griffith, M. Griffith, and S. J. M. Jones, "Text-mining clinically relevant cancer biomarkers for curation into the CIViC database," *Genome Med.*, vol. 11, no. 1, pp. 1–16, Dec. 2019.

[7] J. Labrosse, T. Lam, C. Sebbag, M. Benque, I. Abdennebi, H. Merckelbagh, M. Osdoit, M. Priour, J. Guerin, T. Balezeau, B. Grandal, F. Coussy, A. Bobrie, L. Ferrer, E. Laas, J.-G. Feron, F. Reyal, and A.-S. Hamy, "Text mining in electronic medical records enables quick and efficient identification of pregnancy cases occurring after breast cancer," *JCO Clin. Cancer Informat.*, vol. 3, no. 3, pp. 1–12, Dec. 2019.

[8] F. H. Borsato, F. O. Aluani, and C. H. Morimoto, "A fast and accurate eye tracker using stroboscopic differential lighting," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop (ICCVW)*, Dec. 2015, pp. 110–118.

[9] K. Vo, C. Jaremenko, C. Bohr, H. Neumann, and A. Maier, "Automatic classification and pathological staging of confocal laser endomicroscopic images of the vocal cords," in *Bildverarbeitung für die Medizin*. Berlin, Germany: Springer, 2017, pp. 312–317.

[10] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2414–2423.

[11] C. Matava, E. Pankiv, S. Raisbeck, M. Caldeira, and F. Alam, "A convolutional neural network for real time classification, identification, and labelling of vocal cord and tracheal using laryngoscopy and bronchoscopy video," *J. Med. Syst.*, vol. 44, no. 2, pp. 1–10, Feb. 2020.

[12] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[13] R. Janssens, G. Zeng, and G. Zheng, "Fully automatic segmentation of lumbar vertebrae from CT images using cascaded 3D fully convolutional networks," in *Proc. IEEE 15th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2018, pp. 893–897.

[14] T. Ivanovska, A. Daboul, O. Kalentev, N. Hosten, R. Biffar, H. Völzke, and F. Wörgötter, "A deep cascaded segmentation of obstructive sleep apnea-relevant organs from sagittal spine MRI," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 16, no. 4, pp. 579–588, Apr. 2021.

[15] Y. S. Derdiman and T. Koc, "Deep learning model development with U-Net architecture for glottis segmentation," in *Proc. 29th Signal Process. Commun. Appl. Conf. (SIU)*, Jun. 2021, pp. 1–4.

[16] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[17] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[18] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 91–99.

[19] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[20] T. A. I. T. Alang, T. T. Swee, T. J. Hou, L. K. Meng, S. A. Malik, M. A. Asari, A. Hussein, A. Yaakub, H. Y. Chai, and J. Haron, "Global Canny algorithm based on Canny edge detector framework in magnetic resonance imaging," *Malaysian J. Fundam. Appl. Sci.*, vol. 13, nos. 4–2, pp. 445–451, Dec. 2017.

[21] R. M. Gurav and P. K. Kadbe, "Real time finger tracking and contour detection for gesture recognition using OpenCV," in *Proc. Int. Conf. Ind. Instrum. Control (ICIC)*, May 2015, pp. 974–977.

[22] R. Šifrer, J. A. Rijken, C. R. Leemans, S. E. J. Eerenstein, S. van Weert, J.-J. Hendrickx, E. Bloemena, D. A. Heuveling, and R. N. P. M. Rinkel, "Evaluation of vascular features of vocal cords proposed by the European laryngological society," *Eur. Arch. Oto-Rhino-Laryngol.*, vol. 275, no. 1, pp. 147–151, Jan. 2018.

[23] P. Gorphe, "A contemporary review of evidence for transoral robotic surgery in laryngeal cancer," *Frontiers Oncol.*, vol. 8, p. 121, Apr. 2018.

[24] R. Obid, M. Redlich, and C. Tomeh, "The treatment of laryngeal cancer," *Oral Maxillofacial Surg. Clinics*, vol. 31, no. 1, pp. 1–11, 2019.

[25] J. M. Lancer, D. Syder, A. S. Jones, and A. Boutillier, "Vocal cord nodules: A review," *Clin. Otolaryngol.*, vol. 13, no. 1, pp. 43–51, Feb. 1988.

[26] R. H. Colton, P. Woo, D. W. Brewer, B. Griffin, and J. Casper, "Stroboscopic signs associated with benign lesions of the vocal folds," *J. Voice*, vol. 9, no. 3, pp. 312–325, Sep. 1995.

[27] J. Ren et al., "Automatic recognition of laryngoscopic images using a deep-learning technique," *Laryngoscope*, vol. 130, no. 11, pp. E686–E693, 2020.

[28] O. Kleinsasser, "Pathogenesis of vocal cord polyps," *Ann. Otol., Rhinol. Laryngol.*, vol. 91, no. 4, pp. 378–381, Jul. 1982.

[29] S. A. Azer, "Challenges facing the detection of colonic polyps: What can deep learning do?" *Medicina*, vol. 55, no. 8, p. 473, Aug. 2019.

[30] Y. Sim, M. J. Chung, E. Kotter, S. Yune, M. Kim, S. Do, K. Han, H. Kim, S. Yang, D.-J. Lee, and B. W. Choi, "Deep convolutional neural network–based software improves radiologist detection of malignant lung nodules on chest radiographs," *Radiology*, vol. 294, no. 1, pp. 199–209, Jan. 2020.

[31] Z. Yan, H. Zhang, R. Piramuthu, V. Jagadeesh, D. DeCoste, W. Di, and Y. Yu, "HD-CNN: Hierarchical deep convolutional neural networks for large scale visual recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2740–2748.

[32] S. Toraman, T. B. Alakus, and I. Turkoglu, "Convolutional CapsNet: A novel artificial neural network approach to detect COVID-19 disease from X-ray images using capsule networks," *Chaos, Solitons Fractals*, vol. 140, Nov. 2020, Art. no. 110122.

[33] W. Zhang, P. Tang, and L. Zhao, "Remote sensing image scene classification using CNN-CapsNet," *Remote Sens.*, vol. 11, no. 5, p. 494, Feb. 2019.

[34] D. I. J. Jacob, "Performance evaluation of caps-net based multitask learning architecture for text classification," *J. Artif. Intell.*, vol. 2, no. 1, pp. 1–10, Mar. 2020.
[35] R. Mukhometzianov and J. Carrillo, "CapsNet comparative performance evaluation for image classification," 2018, *arXiv:1805.11195*.
[36] H. Chao, L. Dong, Y. Liu, and B. Lu, "Emotion recognition from multiband EEG signals using CapsNet," *Sensors*, vol. 19, no. 9, p. 2212, 2019.
[37] Y. LeCun. (2015). *LeNet-5, Convolutional Neural Networks*. [Online]. Available: http://yann.lecun.com/exdb/lenet
[38] W. K. Cho and S.-H. Choi, "Comparison of convolutional neural network models for determination of vocal fold normality in laryngoscopic images," *J. Voice*, vol. 36, no. 5, pp. 590–598, Sep. 2022.
[39] H. Qassim, A. Verma, and D. Feinzimer, "Compressed residual-VGG16 CNN model for big data places image recognition," in *Proc. IEEE 8th Annu. Comput. Commun. Workshop Conf. (CCWC)*, Jan. 2018, pp. 169–175.
[40] H. Salem, G. Attiya, and N. El-Fishawy, "Intelligent decision support system for breast cancer diagnosis by gene expression profiles," in *Proc. 33rd Nat. Radio Sci. Conf. (NRSC)*, Feb. 2016, pp. 421–430.
[41] M. Atlam, H. Torkey, H. Salem, and N. El-Fishawy, "A new feature selection method for enhancing cancer diagnosis based on DNA microarray," in *Proc. 37th Nat. Radio Sci. Conf. (NRSC)*, Sep. 2020, pp. 285–295.
[42] N. Davaris, N. Esmaeili, A. Illanes, A. Boese, M. Friebe, and C. Arens, "Use of artificial intelligence (AI) for the intraoperative evaluation of vocal fold leukoplakias," *Laryngo-Rhino-Otologie*, vol. 100, no. 2, p. S38, 2021.
[43] Q. Zhao, Y. He, Y. Wu, D. Huang, Y. Wang, C. Sun, J. Ju, J. Wang, and J. J. Mahr, "Vocal cord lesions classification based on deep convolutional neural network and transfer learning," *Med. Phys.*, vol. 49, no. 1, pp. 432–442, Jan. 2022.
[44] M. L. Daniel, R. Gregory, and R. Vishwanatha, "Uncovering the important acoustic features for detecting vocal fold paralysis with explainable machine learning," *medRxiv*, vol. 11, no. 7, pp. 127–130, 2021.
[45] A. M. Yousef, D. D. Deliyski, S. R. C. Zacharias, A. de Alarcon, R. F. Orlikoff, and M. Naghibolhosseini, "A hybrid machine-learning-based method for analytic representation of the vocal fold edges during connected speech," *Appl. Sci.*, vol. 11, no. 3, p. 1179, Jan. 2021.
[46] W. K. Cho, Y. J. Lee, H. A. Joo, I. S. Jeong, Y. Choi, S. Y. Nam, S. Y. Kim, and S. Choi, "Diagnostic accuracies of laryngeal diseases using a convolutional neural network-based image classification system," *Laryngoscope*, vol. 131, no. 11, pp. 2558–2566, Nov. 2021.
[47] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[48] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, May 2019, pp. 6105–6114.
[49] X. Xia, C. Xu, and B. Nan, "Inception-v3 for flower classification," in *Proc. 2nd Int. Conf. Image, Vis. Comput. (ICIVC)*, Jun. 2017, pp. 783–787.
[50] P. Refaeilzadeh, L. Tang, H. Liu, L. Angeles, and C. D. Scientist, "Cross-validation," *Encyclopedia Database Syst.*, vol. 5, pp. 532–538, Jan. 2020.
[51] H. Salem, K. R. Negm, M. Y. Shams, and O. M. Elzeki, "Recognition of ocular disease based optimized VGG-Net models," in *Medical Informatics and Bioimaging Using Artificial Intelligence*. Cham, Switzerland: Springer, 2022, pp. 93–111.
[52] J. Shin, Y. K. Chang, B. Heung, T. Nguyen-Quang, G. W. Price, and A. Al-Mallahi, "A deep learning approach for RGB image-based powdery mildew disease detection on strawberry leaves," *Comput. Electron. Agricult.*, vol. 183, Apr. 2021, Art. no. 106042.
[53] A. Rath, D. Mishra, G. Panda, and S. C. Satapathy, "Heart disease detection using deep learning methods from imbalanced ECG samples," *Biomed. Signal Process. Control*, vol. 68, Jul. 2021, Art. no. 102820.

CHEN-KUN TSUNG (Member, IEEE) received the B.Sc. and M.Sc. degrees in computer science and information engineering from Dayeh University, Changhua, Taiwan, in 2004 and 2006, respectively, and the Ph.D. degree in computer science and information engineering from the National Chung Cheng University, in 2014. He is currently an Associate Professor of computer science and information engineering at the National Chin-Yi University of Technology, Taichung, Taiwan. His research interests include cloud computing, big data, web-based applications, and combinatorial optimization.

YUNG-AN TSOU is currently an Associate Professor with the School of Medicine, China Medical University Hospital, and a member of the American Head and Neck Society and Taiwan Voice Society. He is also a member of the British Laryngological Association. He works and did researches on phonosurgery for over 15 years. He received his laryngology training fellowship at the Division of Laryngology, University of California, Davis, under Peter Belafsky. He did researches on the AI diagnostic assistant stroboscopy for helping laryngologists to do better clinical teachings and individualized personal voice therapy or phonosurgery.