scientific reports

Check for updates

OPEN

Diagnosis of unilateral vocal fold paralysis using auto-diagnostic deep learning model

Kyoung Ok Yang[1,5], So Young Kim[2,5], Chang Won Kang[1], Jeong Seon Choi[1], Yong Bae Ji[3], Kyung Tae[3], Jun Won Choi[4]✉ & Chang Myeon Song[3]✉

Unilateral vocal fold paralysis (UVFP) is a condition characterized by impaired vocal fold mobility, typically diagnosed using laryngeal videoendoscopy. While deep learning (DL) models using static images have been explored for UVFP detection, they often lack the ability to assess vocal fold dynamics. We developed an auto-diagnostic DL system for UVFP using both image-based and video-based models. Using laryngeal videoendoscopic data from 500 participants, the model was trained and validated on 2639 video clips. The image-based DL model achieved over 98% accuracy for UVFP detection, but demonstrated limited performance in predicting laterality and paralysis type. In contrast, the video-based model achieved comparable accuracy (about 99%) in detecting UVFP, and substantially higher accuracy in predicting laterality and paralysis type, outperforming the image-based model in overall diagnostic utility. These results demonstrate the advantages of incorporating temporal motion cues in video-based analysis and support the use of DL for comprehensive, multi-task assessment of UVFP. This automated approach demonstrates high diagnostic performance and may serve as a complementary tool to assist clinicians in the assessment of UVFP, particularly in enhancing workflow efficiency and supporting multi-dimensional interpretation of laryngeal motion.

Keywords Vocal fold paralysis, Deep learning model, Auto-diagnosis, Laryngeal videoendoscopy, Multi-task learning, Vocal fold movement

Unilateral vocal fold paralysis (UVFP) refers to the impaired movement of one vocal fold (VF), which can significantly impact patients' breathing and phonation. A retrospective study reported that approximately 19–55% of patients with UVFP experienced dysphagia[1]. Therefore, early and accurate diagnosis of UVFP is crucial to prevent complications. The etiology of UVFP is diverse, encompassing malignancies that invade or compress the recurrent laryngeal nerve, idiopathic, iatrogenic, or systemic causes such as thyroid disorders[2]. This etiologic heterogeneity complicates both prediction and prevention of UVFP. While symptoms are often assessed using tools like the voice handicap index and laryngeal electromyography, the primary method for diagnosing UVFP involves evaluating VF movement through laryngeal videostroboscopy or laryngeal videoendoscopy[3]. These examinations rely on physical observations made by laryngologists or endoscopists. However, incorporating automated, objective analysis of dynamic laryngeal motion may complement existing diagnostic methods, such as electromyography, by enhancing consistency and interpretability in videoendoscopic evaluation.

Several studies have applied DL models to predict VF disorders, such as nodules, polyps, cysts, malignancies, and UVFP. However, a review study[4] found that the majority (13 studies, 93%) of these studies exhibited a high risk of bias, primarily due to small sample sizes, limited demographic diversity, and insufficient external validation of the models. Addressing these limitations requires sufficiently large and diverse datasets, as well as task-specific model designs tailored to the characteristics of each VF disorder. In the case of UVFP, a mobility disorder, dynamic assessment of VF movement is critical for accurately determining both the location and type of paralysis. For instance, a DL-based VF disorder classification model (ODL-VFDDC) used high-speed video endoscopy to analyze vocalized and mobile VF regions[5]. Similarly, models that measured glottic opening angles provided useful metrics but could not represent the complexity of VF motion[6]. These findings underscore the

[1]Department of Artificial Intelligence, Hanyang University, Seoul 04763, Republic of Korea. [2]Department of Anatomy and Cell Biology, College of Medicine, Seoul National University, Seoul 03080, Republic of Korea. [3]Department of Otolaryngology-Head and Neck Surgery, Hanyang University College of Medicine, Seoul 04763, Republic of Korea. [4]Department of Electrical and Computer Engineering, Seoul National University, Seoul 08826, Republic of Korea. [5]These authors contributed equally: Kyoung Ok Yang and So Young Kim. ✉email: junwchoi@snu.ac.kr; cmsong@hanyang.ac.kr

need for more comprehensive approaches that use both static and dynamic features to provide more reliable assessment of VF disorders.

Effective management of UVFP hinges on a comprehensive diagnosis that extends beyond simple detection. Clinically, it is crucial not only to identify the presence and laterality of the paralysis but also to characterize the specific resting position of the paralyzed VF (e.g., median, paramedian, or lateral). This positional characterization provides vital insights into the potential functional impairments, such as adductor or abductor insufficiencies, which directly impact voice quality, airway competence, and inform the selection of appropriate therapeutic strategies. Current diagnostic approaches often rely on subjective visual assessment of VF movement during laryngoscopy, highlighting the need for more objective and reliable analytical tools.

To address these multifaceted diagnostic requirements for UVFP, this study introduces an automated deep learning (DL) model that analyzes continuous VF movement from laryngeal videoendoscopy. The model is designed to perform three hierarchically organized diagnostic tasks: first, identifying the presence of UVFP; second, determining the laterality of the paralysis (left or right); and third, classifying the type of UVFP based on the observed position of the paralyzed fold (e.g., lateral, paramedian, or median). To achieve this, the model utilizes a three-dimensional (3D) video processing backbone to extract critical spatio-temporal features from video sequences. A multi-task learning (MTL) strategy[7] is employed, as MTL is particularly beneficial in this context. It allows the model to learn shared representations relevant to all three diagnostic tasks, leverage the inherent relationships between them, and thereby enhance data efficiency and overall diagnostic performance compared to training separate models for each task.

In this study, we aimed to develop a DL-based classification model capable of analyzing laryngeal videoendoscopy recordings to automatically detect UVFP, determine laterality, and categorize paralysis type. To address the multi-faceted nature of UVFP assessment, we employed a MTL approach, which is well suited for simultaneously optimizing related classification tasks while sharing representations across tasks. This framework does not replace clinical judgment but may serve as a complementary tool to support the interpretation of dynamic laryngeal motion, with the potential to enhance diagnostic consistency and assist in clinical decision-making.

Methods

Data acquisition and participant cohort

This study retrospectively utilized laryngeal videoendoscopy data from 500 participants who underwent examinations at Hanyang University Hospital between March 2013 and July 2019. All data were originally recorded for routine clinical diagnostic and treatment purposes. The study was approved by the ethics committee of Hanyang University (IRB HY-2019-09-004), and the requirement for written informed consent was waived by the Institutional Review Board. All analyses complied with the guidelines and regulations of the ethics committee.

Participant classification

Of the 500 participants, 300 were classified into the normal VF group and 200 into the UVFP group. Classification was based on clinical diagnosis by experienced laryngologists as follows:

- **Normal VF Group:** Participants in this group exhibited normal bilateral VF mobility during phonation tasks and had no other laryngeal pathologies affecting VF movement, as assessed during laryngeal videoendoscopy.
- **UVFP Group:** Participants were included in this group if a clinical diagnosis of UVFP was confirmed via laryngeal videoendoscopy, demonstrating impaired or absent movement of one VF. The specific type of UVFP (paramedian, median, or lateral) was determined by consensus of two experienced otolaryngologists based on the resting position of the paralyzed VF. Median position was defined as paralyzed VF in the midline, and lateral position as paralyzed VF at the fully abducted position. Paramedian position was defined as paralyzed VF in between median and lateral position. Location of UVFP was referred to the side of paralyzed VF as right vs. left.

Laryngeal videoendoscopy protocol

Laryngeal videoendoscopies were performed using an EndoSTROB DX system (Xion, Berlin, Germany). The system's standard built-in light source was utilized for illumination. During the examination, patients were instructed to phonate a sustained vowel /i/ (as in "ee") at a comfortable pitch and loudness to facilitate assessment of VF movement. Original video recordings were captured at a resolution of 1920x1080 pixels and a frame rate of 30 FPS, with the files originally saved in AVI format. These recordings were then processed for the purpose of this study. This processing involved converting the videos and extracting relevant segments into 300x300 pixel MP4 files, which were used for the deep learning analysis. While these recordings were made as part of routine clinical care over several years, examinations were conducted by a consistent team of two experienced laryngologists who followed standard departmental clinical protocols. This approach aimed to maintain procedural consistency and obtain a clear, comprehensive view of the glottis for diagnostic purposes.

From these original recordings, relevant segments clearly depicting VF movement during the /i/ phonation task were manually reviewed and extracted by the clinical team to create the video clips used in this study. These clips, with durations ranging from 1 to 17 seconds, were then processed into 300x300 pixel MP4 files for the subsequent deep learning analysis.

Summary of clinical characteristics

A total of 200 participants with UVFP were 89 (44.5%) male and 111 (55.5%) female participants. The mean age was 63.2 years (SD 13.6) for the UVFP group and 61.8 years (SD 13.0) for the normal control group. The cause

of UVFP was idiopathic (32.5%), cancer (32.0%), iatrogenic (30.5%), tuberculosis (4.5%), and inhalation burn (0.5%). The 164 (82.0%) were on the right side of UVFP and the other (18.0%) were on the left side. The types of UVFP were categorized into paramedian, median, and lateral positions based on the location of the paralyzed VF by two experienced otolaryngologists. The patients with paramedian, median, and lateral positions of UVFP were 55 (27.5%), 71 (35.5%), and 74 (37.0%), respectively.

DL diagnostic model

Proposed network architecture

This study aimed to develop a diagnostic network for UVFP using sequential image data obtained from an endoscopic camera. The proposed DL model was trained to detect UVFP by analyzing the abnormal temporal behavior of the VFs caused by paralysis. Leveraging temporal information from image sequences provides a significant advantage over traditional DL-based diagnostic approaches, which rely solely on spatial features extracted from static images. The proposed DL architecture can diagnose UVFP using endoscopic video footage, offering an advanced approach compared to static image analysis methods.

The proposed DL video network architecture comprises a shared video backbone network and three task-specific head networks, each designed for one of the three diagnostic tasks (Fig. 1). The shared backbone extracts spatiotemporal features from the input video data $T \times H \times W$, where T is temporal dimension, H is height, and W is width.

The proposed architecture leverages 3D video backbone networks, which are designed to capture temporal motion representations from video data. Several well-known architectural candidates for the 3D video backbone are considered. These architectures were selected to represent a diverse range of established and state-of-the-art approaches in video recognition, encompassing 3D convolutional neural networks (CNNs), Transformer-based models, and methods combining 2D CNNs with temporal modeling modules. This diversity allows for a comprehensive evaluation of different strategies for capturing spatio-temporal features relevant to UVFP diagnosis. The candidates include Convolution 3D (C3D)[8], Inflated 3D ConvNet (I3D)[9], SlowFast[10], TimesFormer[11], UniFormerV2[12], Temporal Shift Module (TSM)[13], and Temporal Pyramid Network (TPN)[14]. C3D, I3D, and SlowFast belong to the family of 3D convolutional neural networks, while TimesFormer and UniFormerV2 utilize Transformer-based architectures. In contrast, TSM and TPN apply 2D convolutional neural networks to individual image frames and incorporate additional temporal modeling networks to capture temporal relationships. The extracted 3D features are flattened into a single vector and processed through a linear layer.

The feature vector generated by the 3D backbone network is processed by three task-specific classification heads. Each classification head consists of a two-layer multilayer perceptron (MLP) followed by a softmax layer, which outputs the class probabilities for the respective task. Task 1, processed by the first classification head, identifies whether the VF function is normal or UVFP. Task 2, handled by the second head, determines the location of UVFP, classifying cases as normal, left UVFP, or right UVFP. Task 3, addressed by the third head, provides a detailed characterization of UVFP, distinguishing subtypes such as normal, left paramedian UVFP, left median UVFP, left lateral UVFP, right paramedian UVFP, right median UVFP, and right lateral UVFP. These outputs are computed and delivered concurrently, addressing the specific requirements of each diagnostic task.

The entire network is trained to perform all three tasks simultaneously. The total loss function is defined as the sum of three cross-entropy loss terms corresponding to the tasks:

$$L_{total} = \alpha L_{task1} + \beta L_{task2} + \gamma L_{task3} \tag{1}$$

where L_{task1}, L_{task2}, andL_{task3} are the cross-entropy losses[15] for each task, α, β, and γ are the weight parameters assigned to each task. These were set to α =0.2 for Task 1, β =0.3 for Task2, and γ = 0.5 for Task 3,

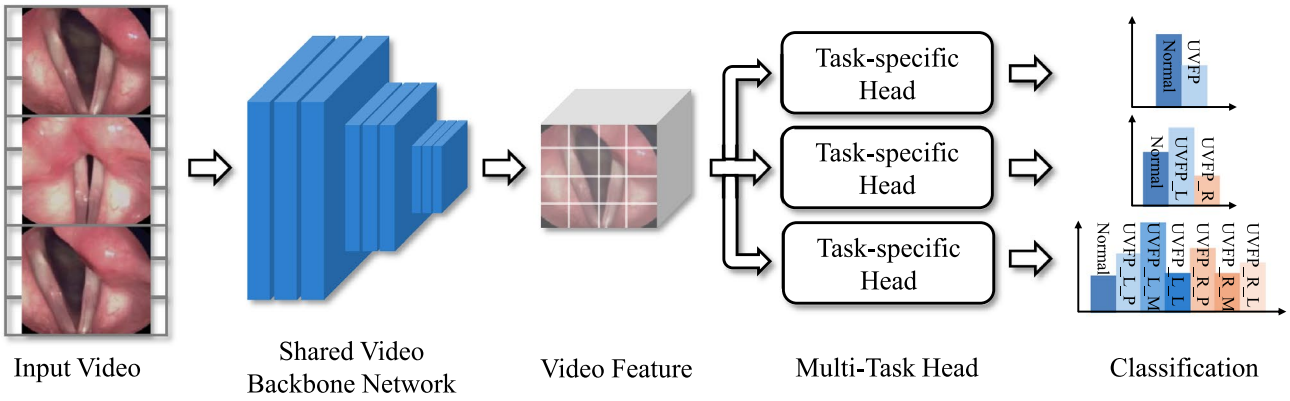


Fig. 1. Conceptual architecture of the proposed network employing a multi-task approach for video diagnostics. The proposed network for video diagnostics utilizes a shared network and multi-task heads for feature extraction from videos, in contrast to the traditional approach that requires a separate video network for each task. The outcomes of multiple tasks are integrated in the final diagnostic stage.

respectively. This weighting strategy reflects an approach where Task 3, being the most granular and complex of the three diagnostic tasks (classifying 7 types versus 2 for Task 1 and 3 for Task 2), was assigned the highest weight. The aim was to ensure that the model dedicated sufficient learning capacity to this more challenging task while still effectively learning the simpler tasks. The specific values were finalized after observing performance trends across multiple experimental runs with different weight combinations to achieve a good balance in performance across all three tasks.

Deep learning model interpretability

In medicine, the readability of DL algorithms used in complex tasks like UVFP assessment is crucial. In particular, understanding how the model arrives at its conclusions is critical for several reasons: (1) to allow clinicians to verify that the model is basing its decisions on clinically relevant features of VF movement rather than confounding artifacts, thereby fostering confidence and facilitating clinical adoption; (2) to enable robust error analysis and guide further model refinement; and (3) to potentially uncover novel visual cues indicative of UVFP that may not be immediately apparent to human observers. Increasing the transparency of the model's decision-making process is key to its responsible and effective integration into clinical workflows.

Gradient-weighted class activation mapping (gradCAM)[16] is a widely used technique that generates heatmaps to highlight the most critical regions within video frames that influence the model's decision. Grad-CAM was selected for this study due to its wide adoption, its applicability to the convolutional neural network (CNN) based architectures explored in our work, and its ability to produce visually intuitive heatmaps that localize important regions without requiring modifications to the model architecture. These heatmaps are particularly useful for video data, as they can illustrate which spatial areas in specific frames contribute most to the model's prediction regarding UVFP characteristics. In this study, Grad-CAM is employed to interpret the decisions made by the proposed DL video model, offering valuable insights into the decision-making process and enhancing its transparency.

Experimental equipment and statistical analysis

In this study, we developed an automated diagnostic program for analyzing VF videos. The performance of the DL video models was evaluated under two scenarios: (1) a single-task learning scenario, where the proposed model was independently trained for each task, and (2) a multi-task learning (MTL) scenario, where the entire network was trained simultaneously on multiple tasks. Various 3D video backbone models were utilized to assess performance across these scenarios.

To ensure a robust evaluation of the generalization ability of our model and to avoid data leakage from individual participants, 5-fold cross-validation was performed at the subject level. Participants were first stratified by gender. Then, within each gender stratum, participants were randomly assigned to one of five groups. Each cross-validation stratum was constructed by combining a group of male participants with a group of female participants, thereby maintaining a consistent sex distribution across folds. In each iteration of cross-validation, four folds were used to train the deep learning models, and the remaining fold was used for testing. This process was repeated five times, with each fold serving as a test set once. The final performance metrics reported in this study are the average of the results obtained from these five test folds.

The diagnostic performance of the DL algorithms was evaluated using Receiver Operating Characteristic (ROC) curve analysis and the Area Under the ROC Curve (AUC), along with Top-1 Accuracy. All statistical analyses were performed using Python version 3.9.2 (Python Software Foundation, Wilmington, Delaware, United States).

Results

Summary of otolaryngology endoscopy video

The clinical dataset consists of 500 videos, categorized into 300 laryngeal videos with normal VF movement and 200 videos of UVFP. The dataset includes 811 clips in the normal group and 914 clips in the UVFP group. Table 1 summarizes the distribution of clips by class. To mitigate potential bias in terms of paralysis of left VF versus right VF in the clinical data, additional video clips were created by horizontally flipping only the original UVFP clips.

These flipped UVFP clips, labeled as "FLIP" in Table 1, effectively balance the dataset regarding laterality. Thus, the total of 2,639 clips was generated from 811 original normal clips, 914 original UVFP clips, and an additional 914 UVFP clips created through flipping (811 + 914 + 914 = 2,639). In total, 2639 clips were generated, with 2111 used for training and 528 for validation.

Performance evaluation of image-based DL model

A preliminary study was conducted to evaluate the performance of a baseline model that utilizes only spatial information from still images. The baseline model was designed to process two images: one depicting the VFs in an open position and the other in a closed position. The architecture of the baseline model is illustrated in Fig. 2. It extracts spatial features from the two images using ResNet18 and ResNet34[17]. ResNet50 was not used due to overfitting issues. Specifically, while ResNet50 achieved high accuracy on the training data, we observed a significantly lower accuracy and an increasing loss on the validation set, indicating poor generalization compared to ResNet18 and ResNet34. The baseline model processes each image separately through two parallel networks based on ResNet architectures to extract spatial features. These extracted features are then concatenated and the combined features are passed through a classification head to produce the probability scores for each specific task.

Table 2 presents the Top-1 accuracy results for each diagnostic task. As shown, both ResNet18 and ResNet34 perform well on Task 1 (detecting the presence of UVFP), achieving accuracies of 99.5% and 98.09%,

		Position		Severity			
	FLIP	Left	Right	Paramedian	Median	Lateral	Number of Clip
Normal	X						811
UVFP	X	✓		✓			167
		✓			✓		186
		✓				✓	358
			✓	✓			57
			✓		✓		91
			✓			✓	55
UVFP	O	✓		✓			57
		✓			✓		91
		✓				✓	55
			✓	✓			167
			✓		✓		186
			✓			✓	358

Table 1. Distribution of video clips by class.

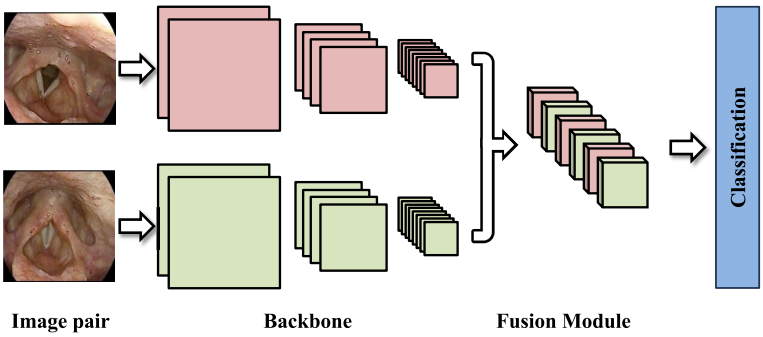


Fig. 2. Architecture of the proposed image UVFP diagnostic network. The model processes paired images of the VFs - one taken when the folds are open and the other when they are closed. Each image is passed through identical image classification networks to extract features. These feature values are then fused to produce a comprehensive diagnostic output using the combined information from both images.

	Top-1 Accuracy (%)		
Model	Task 1	Task 2	Task 3
ResNet18	99.50	65.58	50.30
ResNet34	98.09	65.04	46.92

Table 2. Diagnostic evaluation results obtained when DL models are applied to still images for three tasks: detecting UVFP presence (Task 1), identifying UVFP laterality (Task 2), and classifying UVFP type (laterality and position) (Task 3).

respectively. However, the performance declines significantly for Task 2 (identifying the location of UVFP as left or right), with ResNet18 and ResNet34 achieving accuracies of only 65.58% and 65.04%, respectively. The performance further deteriorates in Task 3, which involves identifying both the location and type of UVFP. Here, the baseline models achieved accuracy of just 50.3% with ResNet18 and 46.92% with ResNet34. These results suggest that while models based on still images can achieve high accuracy for simpler diagnostic tasks (such as UVFP detection in our Task 1), a finding consistent with the successful application of DL to static images in various medical domains for X-ray[18], MRI[19], and CT[20] analysis, they often struggle with more complex tasks requiring detailed spatio-temporal analysis, such as UVFP localization (Task 2) and type classification (Task 3) in our study. This result is consistent with previous research that shows that static analysis of VF conditions has a limitation in capturing dynamic movement patterns[5,6].

Performance evaluation of proposed DL video model

The performance of the DL video models was evaluated under two scenarios: (1) a single-task learning scenario, where the proposed model was independently trained for each task, and (2) a MTL scenario, where the entire

Model		Frame	Top-1 Accuracy (%)		
			Task1	Task2	Task3
Single Task	C3D[8]	16 × 1 × 1	97.67	96.22	92.94
	I3D[9]	8 × 32 × 1	93.31	94.19	83.97
	TSM[13]	8 × 32 × 1	98.26	91.57	90.38
	TPN[14]	8 × 32 × 1	97.67	92.73	88.63
	SlowFast[10]	8 × 32 × 1	98.25	97.67	92.94
	TimeSFormer[11]	8 × 32 × 1	99.42	98.44	94.17
	Uniformer V2[12]	1 × 1 × 8	98.26	97.97	86.88
Ours (Multi-Task)	C3D	16 × 1 × 1	98.83 (+1.16)	98.25 (+2.03)	94.17 (+1.23)
	I3D	8 × 32 × 1	97.67 **(+4.36)**	97.08 **(+2.89)**	93.00 **(+9.03)**
	TSM	8 × 32 × 1	99.42 (+1.16)	99.13 (+7.56)	94.17 (+3.79)
	TPN	8 × 32 × 1	98.25 (+0.58)	97.67 (+4.94)	91.84 (+3.21)
	SlowFast	8 × 32 × 1	98.83 (+0.58)	97.96 (+0.30)	93.88 (+0.94)
	TimeSFormer	8 × 32 × 1	99.42 (–)	99.13 (+0.69)	95.04 (+0.87)
	Uniformer V2	1 × 1 × 8	98.83 (+0.57)	98.54 (+0.57)	94.46 (+7.58)

Table 3. Video recognition model comparison by top-1 accuracy. This presents a comparison of video recognition models based on their top-1 accuracy performance across multiple tasks. The comparison is focused on the models' accuracy performance. Task 1 includes normal and UVFP, Task 2 includes normal, Left UVFP, and Right UVFP, while Task 3 involves normal, Left Paramedian UVFP, Left Median UVFP, Left Lateral UVFP, Right Paramedian UVFP, Right Median UVFP, and Right Lateral UVFP.

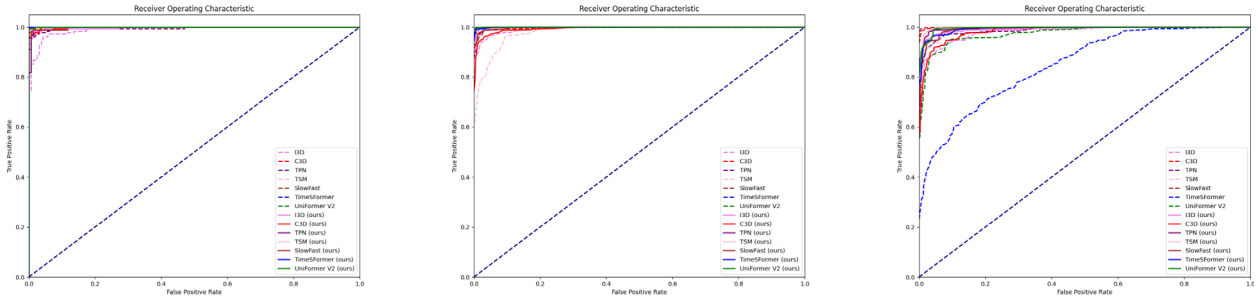
Model		AUC		
		Task1	Task2	Task3
Single task	C3D	0.99	0.99	0.99
	I3D	0.99	0.99	0.98
	TSM	0.99	0.98	0.99
	TPN	0.99	0.99	0.98
	SlowFast	0.99	0.99	0.99
	TimeSFormer	1.00	0.99	0.84
	UniFormer v2	0.99	0.99	0.97
Ours (multi-task)	C3D (Ours)	0.99	0.99	0.98
	I3D (Ours)	0.99	0.99	0.99
	TSM (Ours)	0.99	1.00	0.99
	TPN (Ours)	0.99	0.99	0.99
	SlowFast (Ours)	0.99	0.99	0.99
	TimeSFormer (Ours)	1.00	0.99	0.99
	UniFormer V2 (Ours)	0.99	0.99	0.99

Table 4. Comparison of AUC values for all tasks, depicting the diagnostic performance across Task 1 (Paralysis Presence or Absence), Task 2 (Paralysis Location), and Task 3 (Paralysis Location and Type).

network was trained simultaneously on multiple tasks. Various 3D video backbone models were utilized to assess performance across these scenarios.

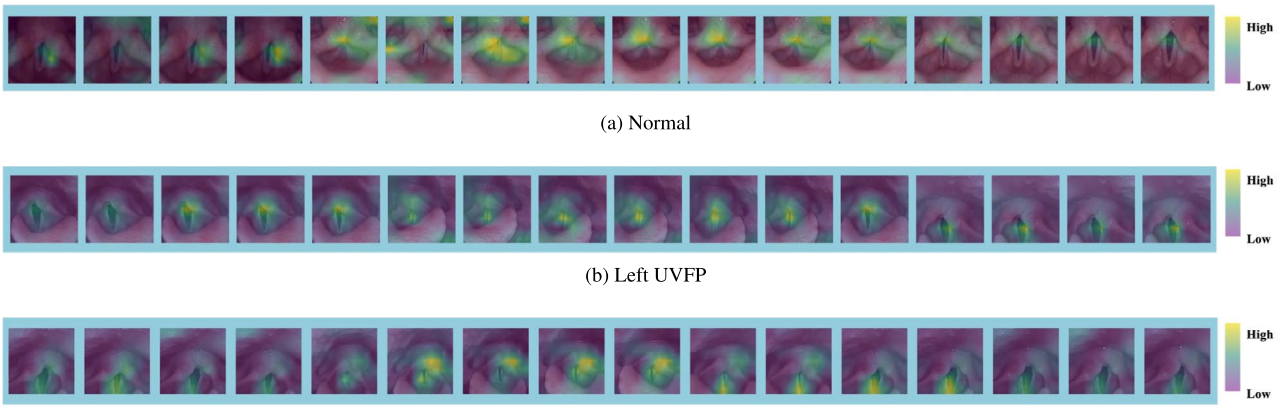
Table 3 presents the Top-1 accuracy of the proposed DL video model, highlighting its significantly superior performance compared to the baseline model (Table 2). These results demonstrate the effectiveness of utilizing sequential video data for diagnosis. With MTL enabled, the Top-1 accuracy exceeds 90% across all tasks. Notably, the TimeSFormer backbone achieves the highest performance, with accuracies of 99.42%, 99.13%, and 95.04% for Task 1, Task 2, and Task 3, respectively. Additionally, the MTL approach consistently outperforms the single-task learning approach for all tasks. The performance gain from MTL is particularly evident in Task 3, where it improves accuracy by 9.03% when using the I3D backbone. However, the performance gain for Task 3 is smaller, at 0.94%, when the SlowFast backbone is employed.

The results of the AUC analysis, along with the corresponding ROC curves, are presented in Table 4 and Figure 3. The proposed deep learning video model achieves an AUC greater than 0.986 when MTL is employed, demonstrating high classification performance on the diagnostic tasks evaluated in this study. While these results are promising, further validation in diverse real-world clinical settings is necessary to ascertain its actual diagnostic utility and reliability in routine clinical practice. In the absence of MTL, there is a significant drop in



(a) Task1 AUC Curve (b) Task2 AUC Curve (c) Task3 AUC Curve

Fig. 3. AUC curve displaying the evaluation results (**a**) Task 1 - presence or absence of paralysis, (**b**) Task 2 - location of paralysis, and (**c**) Task 3 - evaluation of location and severity of paralysis.



(a) Normal

(b) Left UVFP

(c) Right UVFP

Fig. 4. Heatmaps generated by GRAD-CAM. (**a**) Heatmap showing areas of attention for a normal VF. (**b**) Heatmap highlighting areas of attention for left UVFP. (**c**) Heatmap highlighting areas of attention for right UVFP. Yellow regions indicate areas of high attention, while pink regions represent areas of lower attention.

performance for Task 3. The ROC curves generated by the MTL-enhanced approach show high values across all tasks, demonstrating improved and consistent performance, even in complex scenarios such as Task 3.

Interpretability in DL video diagnosis model

Grad-CAM[16] effectively visualizes the network's focus on key regions during VF activity. Figures 4a, b, and c illustrate the heatmaps generated by Grad-CAM on image frames for the normal, left, and right VF cases, respectively. These frames capture the VF cycle, which includes the opening, closing, and reopening phases, allowing the heatmaps to reflect the evolving attention patterns of the proposed video analysis model. To represent these transitions, 16 frames were uniformly sampled from the input video sequence. In the heatmaps, yellow regions indicate areas of high attention, while pink regions represent areas of lower attention.

The figure 4 demonstrates that the proposed DL video model effectively focuses on the VF region, which exhibits motion over time. The Grad-CAM heatmaps show a higher level of model attention, which implies higher activation values, during the closed phase of the VF cycle. While this does not directly measure physiological muscle activity, the model's increased attention to this phase suggests that it may have learned that features that are more prominent or more stable during the closed glottal state are particularly discriminative for the diagnostic tasks. For example, features such as glottal gap, asymmetry, or the appearance of the adducted VFs, which are critical for UVFP diagnosis, may be most clearly evaluated by the model during this phase.

As the cycle progresses, in the illustrative examples shown in Fig. 4, the initial dispersion of attention transitions into a concentrated focus within the VF region. Furthermore, for the presented cases of paralyzed VFs (Fig. 4b and c), distinct attention patterns seemingly associated with arytenoid cartilage activity are observed. For instance, in the specific example of right VF paralysis shown (Fig. 4c), the model's attention is concentrated on the left arytenoid cartilage; similarly, in the illustrative case of left VF paralysis (Fig. 4b), the attention appears to shift to the right arytenoid cartilage. These qualitative observations from representative examples suggest that the model may learn to utilize such contralateral attention patterns as part of its process for localizing paralysis. However, a quantitative analysis across the entire dataset would be needed to confirm

the consistency and generalizability of these specific attention mechanisms and to more definitively understand their contribution to the model's overall diagnostic performance.

Discussion

DL techniques were initially applied to static medical images, such as X-rays[18], MRIs[19], and CT scans[20]. These advancements revolutionized medical imaging by automating interpretation, enabling accurate diagnoses, and efficiently analyzing complex medical data. However, the importance of capturing temporal dynamics and motion patterns in medical data has revealed the limitations of still image-based approaches, particularly in fields where motion is critical to diagnosis, such as laryngology, cardiology, neurology, and orthopedics. To address these limitations and better capture the dynamic nature of physiological processes, video-based deep learning methods have emerged as a natural progression. The transition from image-based DL to video-based DL in medical diagnostics has been driven by the increasing availability of video imaging modalities, including echocardiography, endoscopy, and fluoroscopy. These DL video models analyze sequential data to detect subtle changes, motion patterns, and physiological dynamics over time, thereby enhancing diagnostic accuracy and clinical decision-making. By leveraging temporal information, video-based DL methods hold the potential to improve diagnostic precision, enable earlier disease detection, and enhance patient care across diverse medical specialties.

The diagnosis of laryngeal voice disorders is typically conducted in clinical settings using laryngoscopy, which relies on endoscopic imaging. Although laryngoscopy provides direct visualization of the VFs, its interpretation is often subjective, relying heavily on the clinician's experience and visual judgment. Subtle or early abnormalities may be overlooked, and inter-observer variability can affect diagnostic consistency. To support clinical decision-making and improve accessibility, efforts have been made to develop machine learning and automated diagnostic methods for laryngeal voice disorders. These methods primarily fall into two categories: the analysis of VF images and the analysis of voice characteristics. While voice analysis is effective for identifying general abnormalities, it has significant limitations in diagnosing specific VF disorders with precision. In particular, diagnosing UVFP–a condition characterized by impaired VF movement–requires a diagnostic tool capable of analyzing dynamic movements, which traditional voice-based methods struggle to achieve.

Several prior studies have explored DL-based approaches to diagnosing UVFP[21,22]. However, they relied on static VF photographs and analyzed only a limited number of cases. In contrast, the present study applies DL to actual laryngoscopic video data to assess dynamic VF movement, directly addressing the key limitations of prior static-image-based approaches. By focusing specifically on UVFP and employing a model capable of analyzing sequential motion patterns, our method offers a more clinically relevant and accurate diagnostic aid, particularly for disorders where movement is a primary indicator.

Our model's ability to capture sequential movement patterns addresses the core limitation of prior methods that focused primarily on static VF images or single-dimensional angular measurements. For example, a previous study reported an AI-based system for automated glottis action tracking[23]. Although it included a larger sample size, this study focused on single-dimensional angular measurements, which may overlook complex movement dynamics. Animal studies have also explored VF tracking methodologies[24,25]. These studies demonstrate technical feasibility but do not provide direct clinical validation in humans. While glottic angles provide important diagnostic information, other factors contributing to UVFP must also be considered for comprehensive analysis and physician estimation of VF motion outperformed frame-by-frame analyses based solely on VF angular velocities and angular range of motion[26]. Our study builds on these prior approaches by analyzing multiple aspects of VF motion using video files. By leveraging video data, our method captures dynamic VF motion patterns more comprehensively, enhancing diagnostic accuracy for UVFP. This highlights the importance of holistic motion interpretation and expert-level pattern recognition, which may be missed by frame-based metrics.

In this study, we introduced a DL video model specifically designed for the diagnosis of UVFP. By leveraging video algorithms to extract spatiotemporal features from sequential image frames, the model effectively identified and diagnosed abnormal VF movements caused by paralysis. To the best of our knowledge, this represents the first reported attempt in the literature to apply such an approach.

The proposed DL video model was trained to perform three key tasks: (1) detecting the presence of UVFP, (2) identifying the laterality of paralysis, and (3) pinpointing the location and type of UVFP. During the initial stages of development, we explored an image-based diagnostic approach that used paired closed and open laryngeal images to differentiate between normal VFs and those affected by UVFP. While this approach achieved 90% accuracy for Task 1 (normal vs. UVFP), it failed to deliver satisfactory results for the other two tasks. These findings highlighted a critical limitation of image-based models: their reliance on spatial information alone, which does not capture the dynamic VF movements essential for accurate UVFP diagnosis.

To address these limitations, we developed a video-based DL diagnostic network capable of analyzing temporal patterns in VF movement. By processing sequences of frames, the model provides a more comprehensive and contextual analysis of VF behavior, enabling the detection of subtle movement variations and changes in VF dynamics over time. This approach allowed our model to identify characteristic UVFP irregularities–such as asymmetry, and incomplete VF closure–that are often missed by static image analysis. Consequently, the video-based approach significantly improved diagnostic accuracy across all tasks, demonstrating its potential as a robust tool for UVFP diagnosis.

A key strength of our study is the inclusion of a substantially larger patient cohort compared to previous research. Additionally, it distinguishes itself by leveraging AI to analyze entire video sequences of VF movements rather than relying on static image features such as mucosal waves or VF angles. Importantly, the accuracy of diagnosis using images and videos was evaluated separately, with findings indicating that video-based analysis provides superior precision. The classification of UVFP in this study was conducted by a highly experienced

laryngologist, which significantly enhanced diagnostic accuracy. This collaborative effort between engineering and laryngology underscores the multidisciplinary nature of the research. A review of studies utilizing AI for office-based laryngoscopy revealed that 76% of such studies were conducted by scientists, technologists, engineers, and mathematicians (STEM) specialists, while only 31% involved collaboration with laryngologists[27]. Our study addresses this gap by fostering collaboration between STEM specialists and clinical experts.

Beyond detecting the presence of UVFP, the current model also classified the type of UVFP, demonstrating its versatility. However, some limitations must be considered when interpreting the results. First, the dataset was collected retrospectively from a single institution, which may limit generalizability to other clinical settings. Second, the model has not yet been externally validated across diverse patient populations or device types. Third, while the model showed strong diagnostic performance, we did not include a direct comparison with diagnoses made by expert otolaryngologists, which limits claims about clinical equivalence. Finally, although our model performs multi-task classification, it does not address nuanced clinical presentations such as VF paresis or intermittent dysfunction, which remain challenging even for experienced clinicians.

In addition, the findings of our interpretability analysis using Grad-CAM were primarily based on qualitative observations of representative examples. While these example cases suggest that the model focuses on clinically relevant regions, we did not perform a quantitative analysis, such as using PCA or clustering techniques on heatmaps, on the entire dataset to statistically assess the consistency and variability of these attention patterns. Such an analysis would provide more robust evidence for the generalizability of the observed heatmap features and represents an important direction for future work to further validate the model's decision process.

Data availability

The datasets used and/or analysed during the current study available from the corresponding authors on reasonable request.

Received: 12 March 2025; Accepted: 30 June 2025
Published online: 29 July 2025

References

1. Schiedermayer, B. et al. Prevalence, incidence, and characteristics of dysphagia in those with unilateral vocal fold paralysis. *Laryngoscope* **130**, 2397–2404 (2020).
2. White, M., Meenan, K., Patel, T., Jaworek, A. & Sataloff, R. T. Laboratory evaluation of vocal fold paralysis and paresis. *J. Voice* **31**, 168–174 (2017).
3. Daggumati, S., Panossian, H. & Sataloff, R. T. Vocal fold paresis: Incidence, and the relationship between voice handicap index and laryngeal emg findings. *J. Voice* **33**, 940–944 (2019).
4. Tessler, I. et al. Deep learning in voice analysis for diagnosing vocal cord pathologies: A systematic review. *Eur. Arch. Otorhinolaryngol.* **281**, 863–871 (2024).
5. Sakthivel, S. & Prabhu, V. Optimal deep learning-based vocal fold disorder detection and classification model on high-speed video endoscopy. *J. Healthc. Eng.* **2022**, 4248938 (2022).
6. Adamian, N., Naunheim, M. R. & Jowett, N. An open-source computer vision tool for automated vocal fold tracking from videoendoscopy. *Laryngoscope* **131**, E219–E225 (2021).
7. Caruana, R. Multitask learning. *Mach. Learn.* **28**, 41–75 (1997).
8. Tran, D., Bourdev, L., Fergus, R., Torresani, L. & Paluri, M. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, 4489–4497 (2015).
9. Carreira, J. & Zisserman, A. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 6299–6308 (2017).
10. Feichtenhofer, C., Fan, H., Malik, J. & He, K. Slowfast networks for video recognition. In *ICCV*, 6202–6211 (2019).
11. Bertasius, G., Wang, H. & Torresani, L. Is space-time attention all you need for video understanding? In *ICML*, 4 (2021).
12. Li, K. et al. Uniformerv2: Unlocking the potential of image vits for video understanding. In *ICCV*, 1632–1643 (2023).
13. Lin, J., Gan, C. & Han, S. Tsm: Temporal shift module for efficient video understanding. In *ICCV*, 7083–7093 (2019).
14. Yang, C., Xu, Y., Shi, J., Dai, B. & Zhou, B. Temporal pyramid network for action recognition. In *CVPR*, 591–600 (2020).
15. Zhang, Z. & Sabuncu, M. Generalized cross entropy loss for training deep neural networks with noisy labels. *NeurIPS* **31** (2018).
16. Selvaraju, R. R. *et al.* Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 618–626 (2017).
17. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *CVPR*, 770–778 (2016).
18. Fan, W. et al. A deep-learning-based framework for identifying and localizing multiple abnormalities and assessing cardiomegaly in chest x-ray. *Nat. Commun.* **15**, 1347 (2024).
19. Gassenmaier, S. et al. Deep learning applications in magnetic resonance imaging: has the future become present?. *Diagnostics* **11**, 2181 (2021).
20. Koetzier, L. R. et al. Deep learning image reconstruction for ct: Technical principles and clinical prospects. *Radiology* **306**, e221257 (2023).
21. Voigt, D., Döllinger, M., Yang, A., Eysholdt, U. & Lohscheller, J. Automatic diagnosis of vocal fold paresis by employing phonovibrogram features and machine learning methods. *Comput. Methods Programs Biomed.* **99**, 275–288 (2010).
22. Maniaci, A., Chiesa-Estomba, C. M. & Lechien, J. R. Chatgpt-4 consistency in interpreting laryngeal clinical images of common lesions and disorders. *Otolaryngol. Head Neck Surg.* **171**, 1106–1113 (2024).
23. Wang, T. V. et al. Application of a computer vision tool for automated glottic tracking to vocal fold paralysis patients. *Otolaryngol. Head Neck Surg.* **165**, 556–562 (2021).
24. Haney, M. M., Hamad, A., Leary, E., Bunyak, F. & Lever, T. E. Automated quantification of vocal fold motion in a recurrent laryngeal nerve injury mouse model. *Laryngoscope* **129**, E247–E254 (2019).
25. Pennington-FitzGerald, W. et al. Development and application of automated vocal fold tracking software in a rat surgical model. *Laryngoscope* **134**, 340–346 (2024).
26. Han, J., George, S. S. & Mau, T. Ingredients in the visual perception of hypomobile vocal fold motion impairment. *Laryngoscope* **133**, 866–874 (2023).
27. Yao, P. et al. Applications of artificial intelligence to office laryngoscopy: A scoping review. *Laryngoscope* **132**, 1993–2016 (2022).

Acknowledgements

This work was supported by the research fund of Hanyang University (HY- 202300000003522).

Author contributions

Dr C.M Song and Dr J.W. Choi had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis. Dr K.O. Yang and Dr S.Y. Kim contributed equally to the work as first authors. Dr C.M Song and Dr J.W. Choi contributed equally as corresponding authors. Concept and design: C.M Song, J.W. Choi Acquisition, analysis, or interpretation of data: C.M Song, J.W. Choi, K.O. Yang, S.Y. Kim Drafting of the manuscript: K.O. Yang, S.Y. Kim Critical review of the manuscript for important intellectual content: C.W. Kang, J.S. Choi, Y.B. Ji, K. Tae, J.W. Choi, C.M. Song Statistical analysis: K.O. Yang, S.Y. Kim Obtained funding: C.M. Song Administrative, technical, or material support: C.W. Kang, J.S. Choi, Y.B. Ji, K. Tae, J.W. Choi, C.M Song Supervision: Y.B. Ji, K. Tae. Clinical diagnosis of VF paralysis and categorization of UVFP: C.M.Song, K.Tae

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to J.W.C. or C.M.S.

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.