ENT Updates
https://ojs.ukscip.com/index.php/ENTU

Review

Research Progress in Intelligent Diagnosis of Vocal Fold Lesions Based on Multimodal Deep Learning: A Narrative Review

Ge Gao [1], Kai Zhao [2] and Mingbo Liu [3,*]

[1] Graduate School, Medical School of Chinese PLA, Beijing 100080, China
[2] Department of Otorlaryngology Head and Neck Surgery, Hainan Hospital of Chinese PLA General Hospital, Sanya 572013, China
[3] Department of Otorlaryngology Head and Neck Surgery, Chinese PLA General Hospital, Beijing 100080, China
[*] Correspondence: mingbo666@vip.163.com

Received: 19 December 2025; **Revised:** 9 February 2026; **Accepted:** 3 March 2026; **Published:** 12 March 2026

Abstract: The diagnosis of vocal fold (VF) lesions relies on examinations such as laryngoscopy and voice analysis, which are highly dependent on clinicians' experience. This leads to a relatively higher risk of misdiagnosis and missed diagnosis among junior physicians. In recent years, with the rapid advancement of artificial intelligence (AI), numerous deep learning (DL)-based methods have emerged in this field. Early research primarily focused on single-modality image analysis, such as classifying white light or narrow-band images into benign or malignant lesions using convolutional neural networks (CNNs). However, such methods often fail to fully integrate complementary information from different modalities and fall to meet the clinical demands for multi-classification and risk stratification. Recently, DL and multimodal fusion have gradually become research hotspots. It enables the extraction of complementary multi-category feature information by integrating laryngoscopic images, videos, voice, and clinical text data (e.g., laryngoscopy reports and medical record information), to construct an end-to-end intelligent diagnostic system. This narrative review summarizes the research progress of DL and multimodal fusion in the diagnosis, classification, and severity grading of VF lesions over the past five years (2020–2025). Studies demonstrate that multimodal DL models outperform single-modality models across multiple tasks, which significantly improves the identification and classification accuracy of VF lesions. These models exhibit promising performance. However, DL and multimodal fusion still face numerous challenges, and their clinical translation remains difficult.

Keywords: Vocal Fold; Multimodal; Deep Learning; Intelligent Diagnosis; Laryngoscopic Images; Voice Analysis

1. Introduction

The vocal folds (VFs), situated in the larynx, are responsible for phonation and glottic closure. The common pathological conditions can be categorized as benign lesions (e.g., nodules, cysts, and polyps), precancerous lesions (e.g., leukoplakia, atypical hyperplasia), malignant lesions (e.g., squamous cell carcinoma), and movement disorders (e.g., VF paralysis). These conditions impair voice quality, and can significantly reduce a patient's quality of life. Current diagnosis depends on laryngoscopy, voice analysis, CT, MRI, and etc. Standard laryngoscopy requires inserting a scope into the throat, often causing patient discomfort. It depends on clinical expertise, which may lead to the risk of misdiagnosis and missed diagnosis, particularly for early-stage and micro-lesions.

Advancements in AI, particularly DL, have shown its potential for non-invasive, convenient, and real-time assisted diagnosis. As a subset of machine learning, DL uses multi-layered, end-to-end models to automatically ex-

tract and learn hierarchical features from data, providing a novel approach for the intelligent identification of VF lesions [1]. In 2012, Krizhevsky et al. [2] developed the deep convolutional neural network architecture AlexNet. This landmark achievement surpassed all previous traditional methods and triggered a shift toward DL in the field of computer vision research. He Kaiming et al. [3] designed ResNet, introducing residual connections, which further deepened the capabilities of DL. Ashish Vaswani et al. [4] proposed the Transformer based on the self-attention mechanism, laying the foundation for large language models (LLMs) and multimodal learning. Vision Transformer (ViT) can focus the attention of the network on the main area of the image [5]. Jonathan Ho et al. [6] developed the generative AI model Diffusion Models, advancing AI progress in text-to-image and video generation. In recent years, research on DL in the medical field has become increasingly in-depth, with many encouraging results emerging in the diagnosis of VF lesions.

Multimodal deep learning (MDL), a subfield of DL, refers to neural network models capable of processing and integrating information from multiple distinct modalities (e.g., visual, textual, auditory). It achieves cross-modal information complementarity, correlation, and synergy, thereby enabling disease recognition and diagnosis capabilities that approach or even surpass human performance.

This narrative review aims to summarize the recent 5-year research progress of DL application in the diagnosis of VF lesions, spanning from single-modal (laryngoscopic images, videos, voice signals) to multimodal fusion approaches. Also discusses the attendant challenges and proposes future research directions.

2. Materials and Methods

This narrative review was conducted through an extensive literature search across databases, including PubMed and Web of Science between 2020 and 2025. The search used a combination of keywords including "deep learning", "multimodal", "artificial intelligence", "convolutional neural networks", "vocal fold", and "vocal cord" (AND, OR, and NOT) to refine 191 articles from PubMed and Web of Science. 74 articles are related. Then we manually screened 46 articles that met the criteria.

The inclusion criteria: (1) Studies must explicitly address the diagnosis, classification, or severity assessment of VF lesions, including but not limited to polyps, nodules, cysts, papillomas, leukoplakia, and carcinoma. (2) The study population must consist of human subjects. (3) The core methodology must be based on DL. (4) Only original research articles are included. (5) Publications must be in English and appear in peer-reviewed academic journals.

The exclusion criteria: (1) Studies focusing on non-vocal fold laryngeal pathologies or normal VFs only. (2) Total sample size fewer than 10 patients. (3) Studies employing traditional machine learning or image processing methods. (4) Studies primarily aimed at software or interface development. (5) Studies with ambiguously described methodologies.

3. Deep Learning-Based Diagnosis Based on Laryngoscopy

Laryngoscopic images and videos provide the most direct visual evidence for the diagnosis of VF lesions, encompassing white light laryngoscopy, narrow-band imaging (NBI), videostroboscopy and high-speed videoendoscopy (HSV), etc. Among these, NBI endoscopy can enhance the visualization of the mucosal surface and surrounding microvessels, exhibiting relatively high sensitivity and specificity in the diagnosis of precancerous lesions such as VF leukoplakia [7]. HSV can capture true, cycle-by-cycle VF vibrations independent of sound periodicity, thereby enabling objective analysis of both normal and pathological phonation.

3.1. Static Images

White light laryngoscopy uses broad-spectrum white light to simulate natural vision, providing a comprehensive view of VF morphology, color, and surface contours. It is a routine examination for VF lesions. NBI employs narrow-band light to enhance the contrast between submucosal vessels and surrounding tissues, clearly revealing epithelial abnormalities. It offers higher sensitivity in detecting early-stage cancers and precancerous lesions and assists in delineating the extent of tissue infiltration. CNNs, standing as one of the most effective tools for processing images, are extensively employed in image analysis, recognition and classification of VF endoscopic images. They are widely used in binary classification—such as distinguishing images with or without VFs, identifying healthy versus unhealthy VFs, or differentiating specific pathologies like polyps, leukoplakia, or unilateral vocal fold paralysis

(UVFP). Moving further, CNNs are applied to multi-class classification tasks to distinguish normal VFs from benign and malignant lesions, and to recognize a range of conditions including normality, nodules, polyps, leukoplakia, carcinoma, Reinke's edema, and etc. Additionally, they can be utilized for pathological grading, such as identifying three grades of dysplasia and carcinoma in situ, followed by stratification into high-risk and low-risk groups. CNNs show promising performance in laryngoscopic images classification.

3.1.1. Lesion Detection and Classification

To discriminate abnormal VFs from the normal, Won Ki Cho et al. [8] evaluated four CNN-based DL models (CNN6, VGG16, Inception V3, and Xception). The accuracy of four models was 82.3%, 99.7%, 99.1%, and 83.8%, respectively, in the test set. Based on static image pairs, Kyoung Ok Yang et al. [9] constructed a DL model, employing ResNet18 and ResNet34 as backbone networks, to extract features from paired keyframe images—representing the open and closed states of the VFs. The study extracted images from videos of 500 patients (300 normal, 200 UVFP). The model achieved accuracies of 99.50% and 98.09%, respectively in detecting VF abnormalities. However, solely identifying whether the VFs are normal holds limited practical value in clinical practice. The above two studies are all based on a single-center.

To identify leukoplakia, Mei-Ling Wang et al. [10] developed a multi-instance learning (MIL)-based AI model and validated it on a multicenter white light laryngoscopy dataset. Unlike common DL models, the MIL method aggregates images from the same patient into a "bag" for comprehensive patient-level diagnosis and significantly reduces annotation costs. The results demonstrated that the model exhibited strong diagnostic performance for VF leukoplakia (patient-level AUC improving to 0.869 in the validation set and 0.851 in the external test set). The model showed good agreement with clinical experts and maintained robust real-time diagnostic capability (patient-level AUC of 0.850) in the prospective cohort. However, the study was primarily conducted internally and lacks independent, large-scale, prospective external validation. Furthermore, the introduction of multi-instance learning (MIL) has increased the complexity of the system.

Another multicenter research was conducted to automatically detect laryngeal carcinoma from cases mixed with benign lesions. Peikai Yan et al. [11] used Faster R-CNN, which has a fast detection speed. Faster R-CNN uses a shared convolutional backbone between its region proposal network and detection module to efficiently generate high-quality region proposals while significantly reducing computational costs. The dataset comprised 2,179 laryngoscopy images from 2,179 patients across 6 hospitals, captured using 5 distinct laryngoscopy systems. The model sensitivity, specificity and overall accuracy were 74.16%, 78.59%, and 78.05%, respectively. But the dataset is imbalanced, with benign samples outnumbering malignant samples by more than sixfold. The positive predictive value is only 32.51%, which may lead to unnecessary further examinations.

Since some laryngoscopic images do not contain VFs at all, identifying such non-informative images for subsequent removal is a basic step. Bich Anh Tran et al. [12] compared 5 CNN models (ResNet50V2, MobileNetV2, InceptionV3, DenseNet201, and Xception). Among which, Xception had the highest accuracy of 97.51%. The accuracy of finding no VFs, normal, and abnormal VFs were 98.90%, 97.36%, and 96.26%, respectively and the highest AUC (0.9941). However, this study has no validation cohort, and only 455 images for testing, which may cause the results to be biased. Wellenstei et al. [13] compared three scaled versions of the YOLOv5 CNN architecture (small, medium, and large), which differ primarily in their number of parameters, to classify normal, carcinomas and benign lesions. The combination of YOLOv5s and YOLOv5m provided higher accuracy in most metrics in both the validation set and the test set for carcinoma detection. Sensitivity for classifying normal VF, carcinomas and benign lesions all exceeded 70%. With a detection speed of 63 fps, the ensemble model is suitable for real-time detection. However, the open-source dataset "Laryngoscope8" used in this study lacks full direct histopathological validation.

Another three studies conducted more complex classification tasks. Zhenzhen You et al. [14] compared 6 CNN models (AlexNet, VGG, Google Inception, ResNet, DenseNet, and ViT) for classifying six categories (normal tissues, inflammatory keratosis, mild dysplasia, moderate dysplasia, severe dysplasia, and squamous cell carcinoma). They utilized both white light images and NBI images. GoogLeNet, DenseNet-121, and ResNet-142 performed the highest overall accuracy, with white light images classification reaching up to 95.83% and NBI images classification up to 94.78%. Furthermore, they performed binary predictions on each vocal condition and four-class classification (mild dysplasia and moderate dysplasia are merged to precancerous case, severe dysplasia and squamous cell carcinoma are merged to cancerous case). However, some subjects are severely misclassified due to blurring, un-

derexposure, and artifacts. S. M. N. Nobel et al. [15] innovatively combined EfficientNetV2L (CNN) with LGBM (gradient boosting) for five kinds of VF diseases classification (carcinoma, dysphonia, paresis, polyp, and healthy VFs). Using a training set of over 10,000 images and validation/test sets each exceeding 1,000 images, the model achieved over 97% accuracy across all datasets. For VF segmentation, they integrated UNet (encoder-decoder) with BiGRU (bidirectional gated recurrent unit) to enhance spatial context and long-range dependencies between pixels in laryngeal images. The segmentation model attained 92.55% accuracy on a training set of 18,000 images, with 89.87% and 91.47% accuracy on validation and test sets of 3,000 images each, respectively. However, BiGRU models spatial dependencies in static images rather than true dynamic temporal sequences, and both models are ensemble-based, resulting in high computational cost and long training times. All experiments were conducted on the Zenodo public dataset without external validation. To classify eight kinds of VF lesions, including normal, glottic cancer, granuloma, Reinke's Edema, VF cyst, leukoplakia, nodules and polyps, Ran Wei et al. [16] created a NDF-Net model, combining deep neural networks (DNNs) and ViTs, which preserved both local features and global representations. Based on the Region of Interest (ROI) of the laryngoscope localized by YOLOv5, the overall accuracy of the NDF-Net model was 86.51%, and the average AUC was 0.954. Compared with classification based on the entire images, where F1-Scores based on the ROI are improved except in leukoplakia, especially for granuloma and normal, F1-Scores were 0.958 and 0.921, respectively. In addition to the lack of independent external validation, the dataset used in this study also suffers from class imbalance. For instance, glottic cancer comprises only 22 samples, resulting in poor classification performance of the model for these disease categories.

Another study is worthy of attention. Currently, with the gradual development of generative AI, its application fields are becoming increasingly broad. While traditional studies have exclusively relied on real patient laryngoscopy images, whether from open-source databases or self-collected, Iman Khazrak et al. [17] employed Denoising Diffusion Probabilistic Models (DDPM) to generate 4,180 synthetic white-light laryngoscopy images covering seven types of VF structural pathologies (cysts, granulomas, keratoses, nodules, polyps, Reinke's edema, and sulcus vocalis) as well as normal VFs. Their investigation revealed that training with synthetic data combined with 50% of the team's own collection of 404 images enabled the pre-trained VGG16 model to achieve optimal performance on the multi-class classification task using an independent test set. Additionally, the quality of the synthetic images was evaluated using the Fréchet Inception Distance (FID), which showed that nodules were the most similar to real images (FID = 104.70), while sulcus vocalis had the lowest similarity (FID = 227.31). This study demonstrates that generative AI holds potential value for supplementing and balancing data in AI-driven pathology classification, offering a new approach for expanding databases, although its synthetic capabilities remain currently limited.

3.1.2. Lesion Severity Grading

Clinically, VF leukoplakia of varying pathological grades corresponds to different treatment strategies. For leukoplakia without dysplasia or with mild inflammation, conservative management with regular follow-up is appropriate. At the same time, surgical resection is recommended for dysplastic lesions and malignant tumors. Currently, pathological biopsy—an invasive procedure—remains the only definitive diagnostic method. Therefore, developing DL models to assist in identifying different pathological grades of leukoplakia holds practical clinical value. Zhenzhen You et al. [18] constructed a Siamese network, which consists of two convolutional networks with identical structures and weights, for accurate VF leukoplakia classification (normal tissues, inflammatory keratosis, mild dysplasia, moderate dysplasia, severe dysplasia, and squamous cell carcinoma). This model achieved an accuracy of 97.56%, outperforming six other DL models—AlexNet, VGG Net, Google Inception, ResNet, DenseNet, and ViT—whose accuracies ranged from 65.85% to 92.68%. However, this study also has some limitations. For instance, it only included 45 patients and 32 normal controls, with 44 of the patients being male, indicating potential selection bias. The small sample size increases the risk of overfitting, and the inclusion of only 6 cases of severe dysplasia suggests a spectrum bias.

3.2. Videos

Clinically commonly used dynamic laryngoscopes primarily include conventional electronic laryngoscopes (white light/NBI mode), videostroboscopy, and high-speed videoendoscopy (HSV). The recording frame rate of videostroboscopy is 25–30 fps. It provides information on periodic VF vibrations and occupies minimal storage space. However, it is limited in assessing severely irregular vibrations, such as spasmodic dysphonia. In contrast,

HSV records at frame rates exceeding 2,000 fps, enabling sensitive detection of vibratory details, aperiodic vibrations, and transient abnormalities. But it produces substantially larger file sizes. In this category of research, frames extracted from videostroboscopy and HSV rather than videos are sometimes utilized. Compared to static white-light laryngoscopy images, each frame contains morphological information from different phases of the VF vibration cycle. These frames can identify a localized loss or weakening of the mucosal wave. They can also measure the shape and size of the glottal gap during closure. All these features contribute to the recognition and classification of VF pathologies. However, single-frame images lose the dynamic vibratory characteristics present in videos, which does not align with the real-world clinical setting where videos are observed. This limitation restricts physicians' ability to perform dynamic diagnosis. Therefore, we consider the comparison between a DL model trained on single frames and clinicians to be partially unfair.

3.2.1. Videostroboscopy

DL methods for video analysis can be categorized into two types: one extracts only spatial features from key frames or all frames of the video, aggregates these independent frame features (e.g., by taking maximum or mean values) to output classification results, and essentially does not utilize inter-frame temporal dependencies; the other fully captures the temporal dynamic information of VF movement, conducts joint in-depth spatiotemporal analysis of the video via specialized spatio-temporal modeling architectures (such as 3D convolution, Transformer, and temporal shift module), thereby achieving more comprehensive and accurate diagnostic outcomes.

Similar to static images, applying DL to extract frames from stroboscopic laryngoscopy videos can also be used for VF disease classification. Tsung and Tao [19] developed a CNN-based edge-based VF disease detector (EVC-DD) to recognize nodules, polyps, and cancer. The study utilized a dataset of 1,740 images extracted from videostroboscopy videos of 13 confirmed cases. The images were split into 1,044 (60%) for training, 348 (20%) for validation, and 348 (20%) for testing. Using five-fold cross-validation to ensure robust evaluation, the model demonstrated outstanding performance. In comparative experiments with VGG16, EfficientNet, and InceptionV3, EVC-DD achieved 100% accuracy in detecting VF cancer, matching VGG16. For nodules and polyps, both EVC-DD and VGG16 exceeded 99% overall accuracy, with EVC-DD showing slight advantages in certain metrics and significantly faster training times. By contrast, EfficientNet and InceptionV3 attained lower accuracies of 93.08% and 90.78%, respectively. However, the study is limited by its small sample size and class imbalance (e.g., only 182 polyp images, 10.46%), which may lead to overfitting. Furthermore, the model lacks external validation and is confined to detecting edge-based morphological changes, thus unable to identify lesions on the medial VF surface. For VF edge-related diseases, DL can be directly used to identify abnormal protrusions on the VF edges. For diseases such as VF paralysis, where there are no abnormalities in the VF edges but only movement disorders, Elliana Kirsh DeVore et al. [20] have provided a promising approach. They constructed a DL-based computer vision tool—AGATI—that measures the anterior glottic angle (AGA) across different frames after identifying the VF edges, and further classifies unilateral and bilateral vocal fold paralysis (BVFP) by comparing the AGA values. The model was used to calculate AGA of key frames from videostroboscopy recordings of 70 BVFP, 70 UVFP, and 72 normal patients. The 97th percentile AGA effectively distinguished BVFP from normal VFs with high diagnostic accuracy (AUC 0.92). The study also correlated AGA with patient-reported outcomes and assessed its ability to predict the need for surgical airway intervention in BVFP patients.

As mentioned above, white light or NBI images are predominantly used to identify exophytic or morphologically visible lesions of the VFs, which exhibit relatively distinct spatial features. In contrast, the VF sulcus, characterized by a structural defect in the mucosal layer, often presents with subtle or no obvious morphological abnormalities. Its diagnosis requires analysis of images that capture vibratory information or the application of sequential video-based analysis. Ömer Tarık Kavak et al. [21] utilized a CNN-based model to differentiate sulcus from healthy VFs. They used 11,150 for training, 2,788 for validation, and 3,468 for testing. The model accuracy achieved 98%. In a seven-class classification (healthy, nodule, papilloma, polyp, sulcus, cyst, pseudocyst), the CNN-based model achieved 85% accuracy in the testing set, including 6,319 images. The CNN model outperformed five experienced laryngologists across various metrics, such as accuracy (CNN: 76% vs. Laryngologists' average: −67%), Cohen's Kappa Score (CKS), macro average F1-score, and Matthews Correlation Coefficient (MCC). However, many samples lacked pathological confirmation and were assessed solely by one laryngologist, introducing the risk of label bias and misclassification.

For the second type, Kyoung Ok Yang et al. [9] developed a spatio-temporal DL system for UVFP using both image-based and video-based models. The model consists of a shared spatio-temporal backbone—such as 3D CNN or Transformer architectures—to extract dynamic motion features from videos. This backbone feeds into three task-specific classification heads, which simultaneously perform binary classification (normal vs. UVFP), ternary classification (normal, left UVFP, right UVFP), and seven-class fine-grained classification (normal along with left/right paramedian, median, and lateral UVFP subtypes). In distinguishing normal VFs from UVFP using static images, ResNet18 and ResNet34 achieved accuracies of 99.5% and 98.09%, respectively. However, the accuracy of both models dropped substantially when further tasked with differentiating the side of paralysis (<70%) or localizing the fixation position (<60%). Among DL video models, in a multi-task learning (MTL) scenario, Convolution 3D's (C3D) accuracy exceeded 90% in identifying UVFP, paralysis side and the fixation position. This indicates that video provides a significant advantage over static images, though it requires larger storage space and computational costs.

3.2.2. High-Speed Videoendoscopy

HSV can capture aperiodic and irregular VF vibrations and enables precise quantitative analysis of parameters such as mucosal wave velocity, making it a highly valuable tool for diagnosing complex cases.

Transforming high-dimensional, complex dynamic videos into static images can reduce the computational load on the model. Larsen and Pedersen [22] compared four CNN models (a 5-layer CNN, VGG19, MobileNetV2, and Inception-ResNetV2) for differentiating VF nodules from healthy ones. From 30 high-speed videos, 4,000 balanced frames (2,000 normal and 2,000 with nodules) were extracted manually and split into training (2,800), validation (800), and test (400) sets. On the test set, the models achieved accuracies of 97.75%, 83.5%, 91.5%, and 89.75%, and specificities of 95.5%, 88%, 98%, and 86%, respectively. The lightweight MobileNetV2 achieved the highest precision of 97.7%, identified as the most clinically relevant metric for this benign condition. However, this study is based on single-center, small sample size, and reliance on a single expert for frame selection, which may cause selection bias. Furthermore, it lacks external validation and comparison with clinician classification.

Likewise, dynamic videos can also be transformed into other static images that retain characteristic information, such as Kymographic images and phonovibrograms (PVG), which preserve key motion and vibration features in a more compact, lower-dimensional representation. Kymographic images are static single images condensed from thousands of frames of HSV recordings, containing vibratory information (such as periodicity, symmetry, and mucosal wave propagation) captured by horizontal/vertical scan lines. First, Kumar et al. [23] extracted kymographic images from HSV and constructed a custom 2D CNN containing convolutional to perform binary classification of VF disorders ("healthy vs. disordered" and "healthy vs. muscle tension dysphonia") as well as ternary classification (healthy, functional, and organic). The model achieved accuracies of 94.237%, 94.8%, and 93.1% respectively, avoiding the labor-intensive workload of manually segmenting a large number of video frames. However, the study only included an imbalanced dataset (with only 1–2 cases for some organic disorders). Also, it lacked external validation, and only used a basic custom 2D CNN. Subsequently, B. Panchami and S. P. Kumar [24] used HSV from the Benchmark for Automatic Glottis Segmentation (BAGLS) dataset to generate kymographic images and compared the VF disorder classification performance of five pre-trained models (AlexNet, DenseNet121, InceptionV3, ResNet50v2, Xception). The results showed that DenseNet121 achieved the best overall performance in binary classification (healthy vs. disordered), tertiary classification (healthy, functional, organic), and five-class classification (healthy, muscle tension dysphonia (MTD), atrophy, nodules, and edema), with test set accuracies of 98.8%, 98.81%, and 98.47%, respectively. However, kymographic images do not capture transient changes or three-dimensional dynamics and generally only enable preliminary classification (such as healthy vs. pathological) as demonstrated in the aforementioned studies. In contrast, PVG summarizes the full-cycle vibratory characteristics of VFs through color coding and global mapping, supports quantitative assessment, and allows for the subdivision of VF lesions. Building on the above kymogram-based research, the team [25] further explored PVG images (another visualization modality synthesized from HSV) and integrated multi-scale feature extraction and a channel attention mechanism to construct the PVGNet model. Its performance significantly surpassed that of InceptionResNetV2, VGG19, DenseNet169, and X-Vision Transformer (X-ViT) across binary (healthy vs. pathological), tertiary (healthy, functional, and organic), and multi-class (healthy, MTD, atrophy, nodules, and edema) classification tasks, with both overall accuracy and AUC exceeding 96%.

Spatio-temporal DL plays a crucial role in medical dynamic image analysis. Unlike traditional architectures

that rely solely on spatial information, Attia and Benazza-Benyahia [26] evaluated six spatio-temporal DL models—including CTSNF, 3D CNN, ResNet 3D, R(2+1)D, Video Vision Transformer (ViViT), and TimeSformer—for identifying VF disorders (healthy, polyp, paresis, and Reinke edema). Among them, ViViT and TimeSformer achieved the best performance, with F1-scores of 0.9296 and 0.9357, respectively. The spatio-temporal models improved the F1-score by approximately 31% compared to the purely spatial model EfficientNet-B0. However, this study used fixed-length sliding windows (J = 10 frames) to extract clips, which may not fully capture complete glottal vibration cycles even with overlapping (Q = 9). Additionally, the preceding U-Net LSTM (long short-term memory) segmentation of the glottal area could introduce inaccuracies affecting the classification. The research also relied solely on a single public dataset (Zenodo) without external validation.

4. Deep Learning Analysis Based on Voice Signals

Voice changes can occur early in VF disorders. Currently, in addition to laryngoscopy, voice analysis is widely used in clinical practice as a non-invasive examination that requires only speech collection. DL has demonstrated potential in assisting clinicians with clinical diagnosis. It is hoped that in the future, it can further support large-scale screening and enable the development of lightweight models embedded in portable and accessible devices such as mobile phones to aid self-assessment.

4.1. Voice Collection

Voice collection is typically conducted in hospital consultation rooms and specialized laboratories, where the requirements for equipment are relatively high. To assist primary care physicians in making preliminary diagnoses for patients with voice disorders in primary care settings, Evan C. Compton et al. [27] developed an AI-based voice analysis tool. They prospectively collected 203 voice samples recorded using three models of smartphones and supplemented training with voice samples from the Saarbrucken Voice Database (SVD) for comparison. This neural network model achieved higher accuracy than both general practitioners and otolaryngologists in both binary classification (normal vs. abnormal) and multiclass tasks (laryngitis, unilateral paralysis, normal, adductor spasmodic dysphonia (ADSD), and mass lesions). The study highlights the feasibility of AI-based diagnosis in portable and accessible tools, though the sample size is relatively small, and differences in languages and recording conditions across databases limit the model's generalizability. However, this study only changed the recording equipment while still conducting the process in a standard clinic room, with the acoustic environment being processed. It did not fully simulate primary healthcare conditions.

4.2. Voice Processing

In DL-based VF lesion classification tasks, both Mel-spectrograms and Mel-frequency Cepstral Coefficients (MFCCs) are widely used for processing acoustic signals. The Mel-spectrogram retains complete spectral details and is suitable for more complex VF pathology classification, though it requires the model to possess strong feature selection capabilities. MFCCs are cepstral features obtained by applying Discrete Cosine Transform (DCT) compression to the Mel-spectrogram, primarily preserving the "shape" information of the spectrum (such as formant structures). However, this compression process may lead to the loss of critical details, while offering advantages in computational and storage efficiency. Zhen Chen et al. [28] compared the performance of different acoustic features (MFCCs vs. Mel-spectrograms) in the automated detection of dysphonia. The study included data from 461 self-collected participants and 200 subjects from the SVD, employing a 2D CNN to distinguish between normal and dysphonic voices. The results demonstrated that Mel-spectrograms outperformed all MFCC variants, and neither increasing the dimensionality of MFCCs nor adding dynamic features significantly improved performance. Both internal and external test sets achieved accuracy rates above 90%. This study was the first to systematically compare MFCCs and Mel-spectrograms in a DL model. Although the study utilized Mandarin Chinese and German, it employed sustained vowels (/a/ and /i/). Compared to the internal dataset, where each vowel lasted 2–8 s, the external validation set (SVD) had durations of only 1.5–2.6 s, allowing only one effective segment to be generated per subject.

While most studies rely on MFCCs or Mel-spectrograms to analyze voice features, researchers led by Jaemin Song et al. [29] introduced Octave Frequency Spectrum Energy (OFSE) as a new approach, finding that it can detect

finer details in the low-frequency range of voices affected by laryngeal diseases. In their experiment, OFSE features fed into a CNN model achieved 93.98% accuracy in distinguishing four voice conditions—much higher than MFCC's 70.61%. The study, however, had limitations such as a small dataset, uneven groups in terms of disease and gender, and the use of short 0.5-s audio clips that could miss important diagnostic clues.

MFCC and Mel-spectrograms are linear acoustic features that reflect the frequency-domain energy distribution of a signal. In contrast, the production of human speech exhibits nonlinear and even chaotic characteristics. Deli Fu et al. [30] used phase space reconstruction theory to convert one-dimensional speech signals into high-dimensional trajectory images as input to a CNN, employing a VGG-like CNN, to recognize pathological voice. They. Data augmentation was applied across three databases: Massachusetts Eye and Ear Infirmary Database (MEEI), SVD, and a self-built clinical database, and the model was trained using 5-fold cross-validation. In addition to binary classification (normal vs. pathological), the study further conducted a three-class classification task aimed at distinguishing normal voice, VF paralysis, and VF non-paralysis. The results showed that the three-channel phase-space projection achieved average accuracy rates above 95 % across all databases for binary classification, outperforming the single-channel approach. In the three-class setting, the model attained average recognition rates of 96.04% on the MEEI database and 92.27% on the SVD database. This highlights the importance of multi-directional trajectory information for classification. It is a typical example of combining nonlinear theory with DL. However, the pathological voice data in this study were annotated by experts, leading to subjectivity and a lack of uniform standards, which may impact the model's generalization ability.

Currently, like the mentioned studies above, the vast majority of research extracts features directly from raw speech signals and employs DL models to learn the differences among them for classification. However, Mounira Chaiani et al. [31] innovatively proposed and validated the hypothesis that "pathological voice is intrinsically noisy." Based on this premise, they introduced speech enhancement as a critical preprocessing stage in the voice disorder classification pipeline. Their key technical contribution is the development of the Cepstral Harmonics-to-Noise Ratio (CHRN), a dedicated metric for estimating the intrinsic pathological noise within voice signals. The CHNR-based enhancement process aims to suppress non-discriminative noise components while preserving and clarifying the disease-characteristic features in the spectrogram. These enhanced spectrograms are then input into a novel CNN-LSTM network equipped with a newly proposed Sinusoidal Rectified Unit (SinRU) activation function for classification. Experimental results demonstrated the superiority of this approach, particularly achieving a remarkably high accuracy of 99.34% in the challenging task of three-level dysarthria severity classification, significantly outperforming baseline methods. However, during the enhancement process, there is a risk of mistakenly removing crucial pathological feature signals. Additionally, the underlying hypothesis may not be applicable to all types of voice disorders, such as abnormal pitch or loudness. Furthermore, compared to end-to-end models, this framework requires higher computational costs, potentially leading to slower inference speeds.

4.3. VF Lesion Detection and Classification

DL has demonstrated excellent performance in voice disorder identification and classification based on speech signals compared with other methods, and its diagnostic potential is of significant importance for assisting clinical decision-making.

Hyun-Bum Kim et al. [32] employed one CNN, two machine learning models (support vector machine (SVM), LightGBM) and one artificial neural network (ANN) to classify healthy, laryngeal cancer, VF paralysis, and benign mucosal diseases. The laryngeal cancers included in the study consist of glottic and transglottic types. 363 male participants at a single center were included. The voice signals were converted into MFCCs, and 5-fold cross-validation was applied. The results showed that in binary classification tasks (healthy vs. another three conditions), all models achieved accuracy rates above 80%, while in three-class classification (healthy, laryngeal cancer, and VF paralysis/benign mucosal diseases), accuracy ranged from 74% to 83%, with CNN demonstrating superior performance. However, the four classification accuracies of CNN were only 75%. Overlapping features between benign diseases and healthy voices led to a decline in model accuracy in multi-class tasks. However, the study did not account for differences in female voice characteristics or disease distribution. Additionally, the number of laryngeal cancer cases was limited (only 30 cases). Future work could expand the sample size for laryngeal cancer and further classify cases into supraglottic, glottic, and subglottic subtypes.

Apart from comparing DL models with other types of models, some research has conducted comparisons be-

tween different DL models themselves. Awad, A. et al. [33] compared the performance of four network architectures—multilayer perceptron (MLP), regularized MLP, CNN, and LSTM—combined with three acoustic features (spectrogram, Mel-spectrogram, and MFCCs) for hyperkinetic dysphonia diagnosis. They selected 58 normal voice samples and 70 hyperkinetic dysphonia samples from the public VOice ICar fEDerico II Database (VOICED). The results indicated that the regularized MLP model using Mel-spectrogram features achieved optimal performance. All six metrics—accuracy, precision, recall, F1-score, specificity, and AUC—exceeded 99.8%, which significantly outperformed other comparative models and methods. The sample size of this study is small, but the introduction of regularization, Dropout, and other techniques has mitigated the overfitting issue to a certain extent. But splitting each voice recording into 5 or 10 segments may result in fragments from the same recording appearing in both training and test sets, violating the assumption of data independence.

In addition to using a single DL model, integrating multiple DL models and averaging their outputs is also a strategy to reduce prediction errors and achieve more robust results. Hao-Chun Hu et al. [34] utilized pre-trained CNN models (EfficientNet-B0 to B6, SENet154, Se_resnext101_32x4d, and se_resnet152) for transfer learning. They collected a total of 741 voice samples from two hospitals, with 474 used for training, 119 for validation (no independent external validation set), and 148 for testing. Mix-up data augmentation and oversampling were utilized to address class imbalance. Voice signals were converted into MFCCs. The results showed that in the binary classification task of distinguishing between normal and pathological voices, both accuracy and sensitivity exceeded 95%, with a specificity of 84%. In the five-class disease classification task (normal, adductor spasmodic dysphonia, organic VF lesions, unilateral vocal paralysis, vocal atrophy), both accuracy and sensitivity dropped to approximately 66%, yet still outperformed the four clinical physicians involved in the study. However, the collected dataset suffers from a limited sample size and class imbalance—for example, adductor spasmodic dysphonia was represented by only six cases in the test set.

Currently, the paradigm of employing pretrained DL models coupled with classifiers is the predominant technical approach for solving classification tasks in AI across fields such as image and speech processing. Compared to traditional methods, pretrained DL models obtained through self-supervised learning can automatically acquire general feature representations directly from vast amounts of unlabeled raw data. This significantly reduces the model's dependence on costly and scarce manually annotated data. Furthermore, when applied to specific tasks, such models require only fine-tuning on relatively small-scale, task-specific labeled datasets to quickly and efficiently learn and capture key task-related information, such as pathological features.

Ariel Roitman et al. [35] conducted research on the recognition of VF lesions. They employed HuBERT (Hidden-unit BERT) for transfer learning as a speech feature extractor. The model was pre-trained on human speech in a self-supervised manner and connected to classifier to output the results. It achieved an accuracy of 82% and a precision of 87% in the binary classification task of distinguishing between normal and pathological voices. In the multi-class classification task (e.g., VF scar, benign lesions, psychogenic dysphonia, cancer), it demonstrated notably high performance in identifying laryngeal dystonia, with an accuracy of 93.2% and a specificity of 94.2%. However, the model remains constrained by sample limitations, as it relied solely on the SVD database. Similar to this study, Rahman and Direkoglu [36] also employ a hybrid approach of pre-trained DL models: VGGish, but they utilized ensemble classifier. The VGGish model worked as the core feature extractor to derive 128-dimensional Log-Mel spectrogram embeddings. The ensemble classifier (SVM+LR+MLP) can compensate for the classification limitations of individual models. Furthermore, they performed gender stratification. Whether in binary classification tasks (healthy vs. disordered, hyperfunctional dysphonia vs. VF paresis) or the three-class task (healthy, hyperfunctional dysphonia, VF paresis), the classification accuracy was generally higher for male speakers than for female speakers, which may be related to physiological differences, dataset sample sizes, or the distinctiveness of acoustic features across genders. In binary classification for female speakers, the ensemble classifier performed best (accuracy: 71.54%, F1-score: 71.83), while for male speakers and in the three-class tasks, the best model was predominantly VGGish-SVM.

However, the computational demands of these models are relatively high. In contrast, Xiaoping Xie et al. [37] designed a lightweight voice disease classification model that uses only one convolutional layer and extracts 128-dimensional MFCC as input features. The model has low computational cost and is less prone to overfitting, making it suitable for real-time diagnosis on mobile or portable devices. However, due to its shallow network, its ability to identify subtle and complex pathological features in the voice is limited. The study used recordings from 352

patients, and the model achieved an accuracy of 92% in distinguishing four diseases: spasmodic dysphonia, VF paralysis, vocal nodules, and vocal polyps. This demonstrates that a simple network can also perform classification tasks efficiently, but its capacity for more complex or fine-grained classification is still constrained by network depth.

These studies collectively indicate that we need to select the most suitable DL model based on different classification tasks and target populations.

4.4. VF Lesions Severity Grading

DL models can also be applied to classify the severity of VF disorders based on voice signals. Shuaichi Ma et al. [38] proposed a model named TripleConvNet to assess the severity of UVFP, categorizing it into decompensated, partially compensated, and fully compensated states. To capture dynamic instabilities and subtle transient anomalies in voice signals, first- and second-order differential features were incorporated alongside Mel-spectrograms as input features to enhance classification performance. The model achieved accuracy rates of 95.4%, 84.2%, and 74.3% in binary classification (whether medical intervention is needed), three-class classification (distinguishing the three compensation states), and four-class classification (including healthy voices), respectively. However, it should be noted that there is currently no internationally unified grading guideline for the severity of unilateral VF paralysis. The classification scheme proposed in this study is primarily based on the long-term clinical experience of the research team.

We believe it is necessary to point out that there is currently no unified standard for distinguishing between mild voice disorders and normal voices. To differentiate between them, the following efforts may prove useful:

1. Avoid using simple binary classification, as several previously mentioned studies have focused on multi-class classification of vocal disorders.
2. Employ cross-annotation by multiple experienced speech-language pathologists to reduce subjective bias.
3. Combine multi-dimensional acoustic features, including both linear and non-linear characteristics, such as MFCC and the phase space reconstruction method introduced by Deli Fu et al. [30].
4. Adopt temporally sensitive deep learning models, such as LSTMs used by Mounira Chaiani et al. [31] and Awad, A. et al. [33], to capture long-term dependencies and detect persistent subtle anomalies. Incorporate attention mechanisms to focus on abnormal segments.
5. Utilize ensemble learning methods, as demonstrated in the study by Hao-Chun Hu et al. [34], to reduce errors associated with single-model approaches.
6. Implement multimodal fusion by integrating additional data, such as laryngoscopic images, to enhance judgment accuracy.

5. Research Progress in Multimodal Fusion Diagnostic Models

MDL integrates features from multiple aspects of lesions, making diagnostic evidence more comprehensive. It is currently a major hotspot in the intelligent diagnosis of VF lesions. Multimodal data involved in the diagnosis of VF lesions primarily include: visual modalities (laryngoscopy, HSV, CT, MRI, PET-CT, etc.), text modalities (electronic medical records, various diagnostic reports, etc.), and acoustic modalities (voice signals, electroglottography (EGG), etc.). The key to MDL lies in fusion, which typically occurs at three levels: data-level fusion (early fusion), feature-level fusion (mid-level fusion), and decision-level fusion (late fusion). In the diagnosis of VF lesions, feature-level fusion is the most prevalent approach. It involves extracting high-dimensional features from modalities, followed by concatenation or cross-attention operations in the latent space to achieve complementary information exchange between modalities. **Figure 1** provides an overview of multimodal feature fusion and its application. First, data are collected, such as laryngoscopic images, videos, voice signals, text, and other modalities. Subsequently, DL models are used to process and extract features from every single modality. Then, multimodal fusion is performed, which can occur at three levels: data-level fusion takes place before feature extraction by directly concatenating raw data; feature-level fusion occurs after deep learning models independently extract features, where fusion is implemented at intermediate layers; decision-level fusion merges the outcomes after each modality completes its prediction independently. This enables the identification and classification of VF lesions, grading of pathological severity, and other diagnostic objectives.

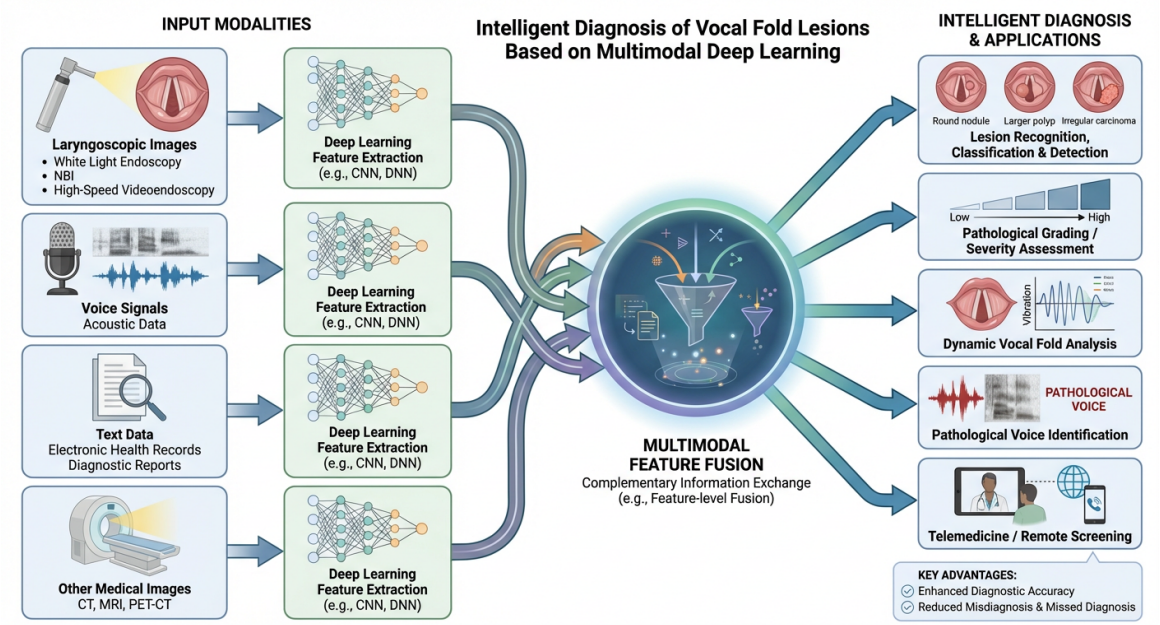

Figure 1. A schematic diagram of the workflow for MDL in vocal fold lesion diagnosis.

5.1. Image-Text Fusion

Zhaohui Jin et al. [39] proposed a multimodal fusion network (VLMF-Net) for early diagnosis of glottic carcinoma. The model integrates laryngoscopic images and clinical text reports, utilizing ViT for image feature extraction and Large Language Model Meta AI (LLaMA3) for textual feature extraction. A Q-Former module from BLIP-2 is introduced to align cross-modal features and reduce inter-modal discrepancies. Through language-guided visual attention, the model focuses on key regions described in the text, achieving feature-level multimodal fusion. On the internal test set, VLMF-Net achieved an accuracy of 77.6%, while maintaining 73.9% on an external independent test set, demonstrating strong generalization capability. The multimodal model significantly outperformed its single-modal counterparts. The study focused on binary classification (early glottic carcinoma vs. dysplasia). Future work should expand to multiclass laryngeal lesions diagnosis to enhance clinical applicability.

5.2. Voice-Text Fusion

Chi-Te Wang et al. [40] integrated voice signals, demographic data (e.g., sex, age), and 26 items of structured medical records (e.g., the onset and presentation of voice symptoms, occupational vocal demand, and smoking status and alcohol consumption) through feature-level and decision-level fusion, employing a DL model to automatically detect glottic neoplasm from benign voice disorders. A split-fusion architecture was adopted, where a backbone feature extraction network processed the voice signals, consisting of five consecutive convolutional blocks, with the introduction of bidirectional gated recurrent units to capture temporal dependencies and an attention mechanism. Non-voice data were handled by a multimodal fusion module. A 10-fold stratified cross-validation was implemented, with data partitioned into training, validation, and testing sets in an 8:1:1 ratio, and predictions were further performed on an independent external dataset (SVD). The results demonstrated that the three-modal fusion outperformed both the single voice modality and the two-modal fusion of voice and demographic features, achieving an accuracy of 82%, which was comparable to the expert assessment accuracy of 83%. Experts exhibited higher specificity (86%–98%) but lower sensitivity (43%–82%), potentially leading to missed diagnoses, whereas AI showed higher sensitivity (48%–78%) but lower specificity (43%–82%). This suggests a future potential for a collaborative diagnostic model of "AI preliminary screening + expert review." However, the main limitations of the study remain the small dataset size and narrow disease spectrum—particularly, only 60 cases of glottic neoplasm were included in the collected cohort, and merely 23 cases were used for external validation, which constrain the model's generalizability.

5.3. Vision Fusion

Although both white light laryngoscopy images and NBI are image-based modalities, they provide different types of detailed features. Therefore, they can still be fused, typically at the feature level. Cheng-Wei Tie et al. [41] developed an AI model based on MIL, which integrates white-light imaging (WLI) and NBI through feature-level fusion. In this approach, the DL model first performs automatic segmentation of lesion regions. The model achieved patient-level AUCs of 0.868 and 0.884 on the internal and external validation sets, respectively, and reached an AUC of 0.825 in the prospective video validation. Also, multimodal fusion-based diagnosis outperformed diagnosis using either single modality alone. Furthermore, the model's performance was significantly superior to that of 12 laryngologists. With AI assistance, the accuracy of junior laryngologists improved from 0.717 to 0.841. AI assistance also improved junior endoscopists' F1-score for high-risk lesion identification from 0.596 to 0.750, and for malignant lesions from 0.737 to 0.845, approaching the level of senior laryngologists. But the study was retrospective in design, and future real-time performance evaluation is needed. Additionally, it lacks interpretability and has high computational demands.

6. Challenges and Prospects

Deep learning and multimodal fusion have demonstrated considerable potential, offering faster decision-making compared to human capabilities. We hope that in the future, such technologies can serve as auxiliary tools to help clinicians make rapid diagnoses and improve workflow efficiency. They may also assist in reducing unnecessary invasive procedures. Furthermore, we even envision that real-time diagnostic capabilities could support junior physicians in quickly delineating lesion boundaries during surgeries, thereby shortening operative time. However, there remain numerous significant issues and challenges that cannot be overlooked.

6.1. Small Sample Size and Heterogeneity

Many current studies suffer from limited sample sizes, with particularly few cases for certain lesion types [7,12] or rare diseases, leading to class imbalance. Insufficient training data adversely affects model classification performance. Additionally, multiple frames or segments extracted from a single patient's laryngoscopic images or voice recordings may appear across training, validation, and test sets, compromising result authenticity. Although data augmentation techniques (e.g., center crop, random rotation, random resized crop, random affine, random horizontal flip) are often applied [7, 14], a study suggests these methods do not consistently improve model performance [42]. Moreover, self-collected or publicly available databases (e.g., "Laryngoscope8") may lack histopathological confirmation, especially for normal VFs or certain benign lesions, relying instead on expert annotations that can be imprecise. Variations in dataset partitioning also lead to inconsistent evaluation metrics [12], with no unified standard currently established.

Furthermore, many studies are single-center, using data from a single laryngoscopy system or video source [7, 12, 14, 18, 43], while multicenter studies involving multiple devices remain scarce. Most rely solely on WLI, though incorporating NBI would better differentiate malignant lesions. Model generalizability is often insufficient due to heterogeneity in laryngoscopy equipment, parameters, resolution, and variations in recording conditions (e.g., background noise, microphone quality, distance, and speaking angle) [44]. These factors degrade model performance and, given the practical difficulty of standardizing such variables, maintaining accuracy across diverse conditions increases model complexity, computational demands, storage requirements, and costs.

Many studies lack external validation and comparisons with clinicians, relying solely on retrospective designs without prospective validation [9]. Most research uses high-quality laryngoscopic images or videos, yet in real-world scenarios, factors such as glare, blur, and airway secretions reduce model diagnostic capability [13,41]. Video-based detection is more challenging than static images under optimal conditions, and real-time diagnostic accuracy remains difficult to guarantee.

6.2. Lack of Model Interpretability and Clinical Trust

The black-box nature of deep learning models hinders clinical translation. Some studies have introduced Grad-CAM to visualize regions of interest [9, 12, 16, 45], while others employ Occlusion Sensitivity to generate interpretability heatmaps and average explainability maps to reveal inter-class "differentiability" [46]. SHAP analysis is

also used to quantify feature contributions, enhancing interpretability [47].

6.3. High Computational Load and Cost

Large convolutional neural networks (e.g., VGG, ResNet) with deep layers and numerous parameters require extended training and inference times, hindering real-time diagnosis. Models incorporating self-attention mechanisms further increase computational complexity, demanding greater memory and even specialized hardware. Processing video data with temporal features using recurrent neural networks (e.g., LSTM) also escalates computational costs. MDL, which relies on cross-modal attention mechanisms, adds structural complexity and computational overhead [40,41].

In resource-limited settings, lightweight CNNs are preferable. When using Vision Transformers or large CNNs, techniques such as model compression, quantization, or distributed training can improve efficiency. For multimodal approaches, early feature fusion or lightweight cross-modal attention mechanisms may reduce computational burden.

6.4. Challenges in Clinical Translation

Numerous issues remain to be addressed at the ethical level. AI diagnosis is typically trained on large-sample data, which may lead to lower diagnostic accuracy for minority groups, such as specific ethnicities or patients with rare diseases, thereby triggering health inequities. As previously mentioned, its decision-making process lacks transparency, resulting in distrust from both doctors and patients, which determines its clinical value. AI can influence human judgment, leading to decision-making uncertainty, and over-reliance may diminish doctors' active decision-making roles, weakening their clinical authority. It is also an issue whether patients have the right to refuse AI involvement in their diagnosis and treatment.

At the legal level, there is currently a lack of relevant laws and regulations. It remains unclear whether AI diagnosis and treatment fall under product liability law or service negligence law. In cases of diagnostic errors or even medical accidents, responsibility allocation among AI companies, doctors, or medical institutions is undefined, as are the specific proportions of liability. The risks of data leakage from AI diagnosis and treatment are present, yet clear limits on data access and correction rights have not been established. With AI integrated into clinical practice, it is necessary to consider whether medical practitioners' standards need redefinition, as well as determining the ownership of patents or copyrights for AI-generated reports.

Regarding regulation, after AI diagnosis is incorporated into clinical use, challenges arise in monitoring performance, assessing algorithm drift, and managing subsequent system updates and maintenance. What's more, model maintenance and updates increase healthcare costs. Traditional regulatory approval processes may struggle to keep pace with the iterative updates of AI diagnosis. In summary, AI must undergo rigorous ethical review, legal approval and strict regulatory scrutiny before clinical application, and there is still a long way to go.

7. Conclusions

Deep learning and multimodal approaches have demonstrated promising potential in the diagnosis of VF lesions, achieving notable accuracy both in the classification of common diseases and in grading lesion severity. However, current research predominantly consists of retrospective studies conducted in single-center settings with small sample sizes and uniform data collection systems. Future research should focus on multicenter, large-scale prospective studies for further validation. Although artificial intelligence holds the promise of transforming future healthcare paradigms, there remains a long road ahead before clinical translation can be fully realized. For instance, it is important to enhance model interpretability to increase clinical trust. Further efforts should also aim to optimize model lightweighting to enable deployment on common devices. We hope that with ongoing progress in multimodal deep learning, it can serve as a valuable assistant in clinical practice, advancing diagnostic precision and expanding the accessibility of medical services in a positive and impactful manner.

Author Contributions

G.G.: Planning, designing, literature survey, evaluation of literature search, writing; K.Z.: supervision; M.L.: funding acquisition. All authors have read and agreed to the published version of the manuscript.

Funding

This work received no external funding.

Institutional Review Board Statement

Not applicable because this is a review paper.

Informed Consent Statement

There was no need to take informed consent because this is a review paper.

Data Availability Statement

All data for this study is presented in this paper.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. LeCun, Y.; Bengio, Y.; Hinton, G. Deep Learning. *Nature* **2015**, *521*, 436–444.
2. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. *Commun. ACM* **2017**, *60*, 84–90.
3. He, K.; Zhang, X.; Ren, S.; et al. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
4. Vaswani, A.; Shazeer, N.; Parmar, N.; et al. Attention Is All You Need. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
5. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; et al. An Image Is Worth 16×16 Words: Transformers for Image Recognition at Scale. *arXiv preprint* **2021**, *arXiv:2010.11929*.
6. Ho, J.; Jain, A.; Abbeel, P.; et al. Denoising Diffusion Probabilistic Models. *arXiv preprint* **2020**, *arXiv:2006.11239*.
7. Zhao, Q.; He, Y.; Wu, Y.; et al. Vocal Cord Lesions Classification Based on Deep Convolutional Neural Network and Transfer Learning. *Med. Phys.* **2022**, *49*, 432–442.
8. Cho, W.K.; Choi, S.-H. Comparison of Convolutional Neural Network Models for Determination of Vocal Fold Normality in Laryngoscopic Images. *J. Voice* **2022**, *36*, 590–598.
9. Yang, K.O.; Kim, S.Y.; Kang, C.W.; et al. Diagnosis of Unilateral Vocal Fold Paralysis Using Auto-Diagnostic Deep Learning Model. *Sci. Rep.* **2025**, *15*, 27635.
10. Wang, M.-L.; Tie, C.-W.; Wang, J.-H.; et al. Multi-Instance Learning Based Artificial Intelligence Model to Assist Vocal Fold Leukoplakia Diagnosis: A Multicentre Diagnostic Study. *Am. J. Otolaryngol.* **2024**, *45*, 104342.
11. Yan, P.; Li, S.; Zhou, Z.; et al. Automated Detection of Glottic Laryngeal Carcinoma in Laryngoscopic Images from a Multicentre Database Using a Convolutional Neural Network. *Clin. Otolaryngol.* **2023**, *48*, 436–441.
12. Tran, B.A.; Dao, T.T.P.; Dung, H.D.Q.; et al. Support of Deep Learning to Classify Vocal Fold Images in Flexible Laryngoscopy. *Am. J. Otolaryngol.* **2023**, *44*, 103800.
13. Wellenstein, D.J.; Woodburn, J.; Marres, H.A.M.; et al. Detection of Laryngeal Carcinoma During Endoscopy Using Artificial Intelligence. *Head Neck* **2023**, *45*, 2217–2226.
14. You, Z.; Han, B.; Shi, Z.; et al. Vocal Cord Leukoplakia Classification Using Deep Learning Models in White Light and Narrow Band Imaging Endoscopy Images. *Head Neck* **2023**, *45*, 3129–3145.
15. Nobel, S.M.N.; Rahman Swapno, S.M.M.; Islam, M.R.; et al. A Machine Learning Approach for Vocal Fold Segmentation and Disorder Classification Based on Ensemble Method. *Sci. Rep.* **2024**, *14*, 14435.
16. Wei, R.; Liang, Y.; Geng, L.; et al. A Non-Local Dual-Stream Fusion Network for Laryngoscope Recognition. *Am. J. Otolaryngol.* **2025**, *46*, 104565.
17. Khazrak, I.; Zainaee, S.; Rezaee, M.M.; et al. Feasibility of Improving Vocal Fold Pathology Image Classification

with Synthetic Images Generated by DDPM-Based GenAI: A Pilot Study. *Eur. Arch. Otorhinolaryngol.* **2025**, *282*, 4139–4153.
18. You, Z.; Han, B.; Shi, Z.; et al. Vocal Cord Leukoplakia Classification Using Siamese Network under Small Samples of White Light Endoscopy Images. *Otolaryngol. Head Neck Surg.* **2024**, *170*, 1099–1108.
19. Tsung, C.K.; Tso, Y.A. Recognizing Edge-Based Diseases of Vocal Cords by Using Convolutional Neural Networks. *IEEE Access* **2022**, *10*, 120383–120397.
20. DeVore, E.K.; Adamian, N.; Jowett, N.; et al. Predictive Outcomes of Deep Learning Measurement of the Anterior Glottic Angle in Bilateral Vocal Fold Immobility. *Laryngoscope* **2023**, *133*, 2285–2291.
21. Kavak, Ö.T.; Gündüz, Ş.; Vural, C.; et al. Artificial Intelligence Based Diagnosis of Sulcus: Assessment of Videostroboscopy via Deep Learning. *Eur. Arch. Otorhinolaryngol.* **2024**, *281*, 6083–6091.
22. Larsen, C.F.; Pedersen, M. Comparison of Convolutional Neural Networks for Classification of Vocal Fold Nodules from High-Speed Video Images. *Eur. Arch. Otorhinolaryngol.* **2023**, *280*, 2365–2371.
23. Kumar, S.P.; Narayanan, N.; Ramachandran, J.; et al. Convolutional Neural Network for Voice Disorders Classification Using Kymograms. *Biomed. Signal Process. Control* **2023**, *86*, 105159.
24. Panchami, B.; Kumar, S.P. Comparison of Deep Learning Models for Voice Disorder Classification Using Kymographic Images. *J. Voice* **2025**, *in press*.
25. Panchami, B.; Kumar, S.P. Comparison of Deep Learning Models for Voice Disorder Classification Using Phonovibrographic Images. *Image Anal. Stereol.* **2025**, *44*, 183–196.
26. Attia, D.; Benazza-Benyahia, A. Recognizing of Vocal Fold Disorders from High Speed Video: Use of Spatio-Temporal Deep Neural Networks. *Int. J. Imaging Syst. Technol.* **2025**, *35*, e70170.
27. Compton, E.C.; Cruz, T.; Andreassen, M.; et al. Developing an Artificial Intelligence Tool to Predict Vocal Cord Pathology in Primary Care Settings. *Laryngoscope* **2023**, *133*, 1952–1960.
28. Chen, Z.; Zhu, P.; Qiu, W.; et al. Deep Learning in Automatic Detection of Dysphonia: Comparing Acoustic Features and Developing a Generalizable Framework. *Int. J. Lang. Commun. Disord.* **2023**, *58*, 279–294.
29. Song, J.; Kim, H.; Lee, Y.O. Laryngeal Disease Classification Using Voice Data: Octave-Band vs. Mel-Frequency Filters. *Heliyon* **2024**, *10*, e40748.
30. Fu, D.; Zhang, X.; Chen, D.; et al. Pathological Voice Detection Based on Phase Reconstitution and Convolutional Neural Network. *J. Voice* **2025**, *39*, 353–364.
31. Chaiani, M.; Selouani, S.A.; Boudraa, M.; et al. Voice Disorder Classification Using Speech Enhancement and Deep Learning Models. *Biocybern. Biomed. Eng.* **2022**, *42*, 463–480.
32. Kim, H.-B.; Song, J.; Park, S.; et al. Classification of Laryngeal Diseases Including Laryngeal Cancer, Benign Mucosal Disease, and Vocal Cord Paralysis by Artificial Intelligence Using Voice Analysis. *Sci. Rep.* **2024**, *14*, 9297.
33. Awad, A.; Eldosoky, M.A.A.; Soliman, A.M.; et al. Automatic Diagnosis of Hyperkinetic Dysphonia from Speech Recordings Based on Deep Learning Approaches. *Eng. Res. Express* **2025**, *7*, 035263.
34. Hu, H.-C.; Chang, S.-Y.; Wang, C.-H.; et al. Deep Learning Application for Vocal Fold Disease Prediction through Voice Recognition: Preliminary Development Study. *J. Med. Internet Res.* **2021**, *23*, e25247.
35. Roitman, A.; Edelstain, Y.; Katzir, C.; et al. Harnessing Machine Learning in Diagnosing Complex Hoarseness Cases. *Am. J. Otolaryngol.* **2025**, *46*, 104533.
36. Rahman, M.U.; Direkoglu, C. A Hybrid Approach for Binary and Multi-Class Classification of Voice Disorders Using a Pre-Trained Model and Ensemble Classifiers. *BMC Med. Inform. Decis. Mak.* **2025**, *25*, 177.
37. Xie, X.; Cai, H.; Li, C.; et al. A Voice Disease Detection Method Based on MFCCs and Shallow CNN. *J. Voice* **2026**, *40*, 524.e1–524.e11.
38. Ma, S.; Liao, W.; Zhang, Y.; et al. Research on Automatic Assessment of the Severity of Unilateral Vocal Cord Paralysis Based on Mel-Spectrogram and Convolutional Neural Networks. *Biomed. Eng. Online* **2025**, *24*, 76.
39. Jin, Z.; Shuai, Y.; Li, Y.; et al. A Vision-Language-Guided Multimodal Fusion Network for Glottic Carcinoma Early Diagnosis: Model Development and Validation Study. *JMIR Med. Inform.* **2025**, *13*, e74902.
40. Wang, C.-T.; Chen, T.-M.; Lee, N.-T.; et al. AI Detection of Glottic Neoplasm Using Voice Signals, Demographics, and Structured Medical Records. *Laryngoscope* **2024**, *134*, 4585–4592.
41. Tie, C.-W.; Li, D.-Y.; Zhu, J.-Q.; et al. Multi-Instance Learning for Vocal Fold Leukoplakia Diagnosis Using White Light and Narrow-Band Imaging: A Multicenter Study. *Laryngoscope* **2024**, *134*, 4321–4328.
42. Yousef, A.M.; Deliyski, D.D.; Zacharias, S.R.C.; et al. Detection of Vocal Fold Image Obstructions in High-Speed Videoendoscopy during Connected Speech in Adductor Spasmodic Dysphonia: A Convolutional Neural Networks Approach. *J. Voice* **2024**, *38*, 951–962.
43. Xiong, M.; Luo, J.-W.; Ren, J.; et al. Applying Deep Learning with Convolutional Neural Networks to Laryngo-

scopic Imaging for Automated Segmentation and Classification of Vocal Cord Leukoplakia. *Ear Nose Throat J.* **2024**, *in press*.
44. Majeed, T.; Assad, A. Advances in Deep Learning for Head and Neck Cancer: Datasets and Applied Methods. *ENT Updates* **2025**, *15*, 1–26.
45. Yao, P.; Witte, D.; Gimonet, H.; et al. Automatic Classification of Informative Laryngoscopic Images Using Deep Learning. *Laryngoscope Investig. Otolaryngol.* **2022**, *7*, 460–466.
46. Özcan, F. Differentiability of Voice Disorders through Explainable AI. *Sci. Rep.* **2025**, *15*, 18250.
47. Ma, K.; Wang, Y.; Zhou, Y.; et al. Acoustic Signatures of Organic Lesions and the Role of Artificial Intelligence in Voice Disorder Diagnostics. *Digit. Health* **2025**, *11*, 20552076251376264.