Nature and Science of Sleep

Elevate Series

Open Access Full Text Article

ORIGINAL RESEARCH

Deep Learning-Based Quantification of Adenoid Hypertrophy and Its Correlation with Apnea-Hypopnea Index in Pediatric Obstructive Sleep Apnea

Jie Cai[1,*], Tianyu Xiu[2,*], Yuliang Song[1], Xuwei Fan[3], Jianghao Wu[1], Aikebaier Tuohuti[1], Yifan Hu[1], Xiong Chen[1,4]

[1]Department of Otorhinolaryngology, Head and Neck Surgery, Zhongnan Hospital of Wuhan University, Wuhan, 430000, People's Republic of China; [2]State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, 430079, People's Republic of China; [3]School of Informatics, Xiamen University, Xiamen, 361000, People's Republic of China; [4]Sleep Medicine Center, Zhongnan Hospital of Wuhan University, Wuhan, People's Republic of China

*These authors contributed equally to this work

Correspondence: Xiong Chen; Tianyu Xiu, Email zn_chenxiong@whu.edu.cn; xiutianyu@whu.edu.cn

Purpose: This study aims to develop a deep learning methodology for quantitative assessing adenoid hypertrophy in nasopharyngoscopy images and to investigate its correlation with the apnea-hypopnea index (AHI) in pediatric patients with obstructive sleep apnea (OSA).

Patients and Methods: A total of 1642 nasopharyngoscopy images were collected from pediatric patients aged 3 to 12 years. After excluding images with obscured secretions, incomplete adenoid exposure, 1500 images were retained for analysis. The adenoid-to-nasopharyngeal (A/N) ratio was manually annotated by two experienced otolaryngologists using MATLAB's imfreehand tool. Inter-annotator agreement was assessed using the Mann–Whitney *U*-test. Deep learning segmentation models were developed with the MMSegmentation framework, incorporating transfer learning and ensemble learning techniques. Model performance was evaluated using precision, recall, mean intersection over union (MIoU), overall accuracy, Cohen's Kappa, confusion matrices, and receiver operating characteristic (ROC) curves. The correlation between the A/N ratio and AHI, derived from polysomnography, was analyzed to evaluate clinical relevance.

Results: Manual evaluation of adenoid hypertrophy by otolaryngologists (p=0.8507) and MATLAB calibration (p=0.679) demonstrated high consistency, with no significant differences. Among the deep learning models, the ensemble learning-based SUMNet outperformed others, achieving the highest precision (0.9616), MIoU (0.8046), overall accuracy (0.9182), and Kappa (0.87). SUMNet also exhibited superior consistency in classifying adenoid sizes. ROC analysis revealed that SUMNet (AUC=0.85) outperformed expert evaluations (AUC=0.74). A strong positive correlation was observed between the A/N ratio and AHI, with the correlation coefficients for SUMNet-derived ratios ranging from r=0.9052 (tonsils size+1) to r=0.4452 (tonsils size+3) and for expert-derived ratios ranging from r=0.4590 (tonsils size+1) to r=0.2681 (tonsils size+3).

Conclusion: This study introduces a precise and reliable deep learning-based method for quantifying adenoid hypertrophy and addresses the challenge posed limited sample sizes in deep learning applications. The significant correlation between adenoid hypertrophy and AHI underscores the clinical utility of this method in pediatric OSA diagnosis.

Keywords: adenoid hypertrophy, obstructive sleep apnea (OSA), deep learning, transfer learning, ensemble learning

Introduction

Adenoid hypertrophy (AH), the enlargement of the adenoids located in the upper throat, is a common cause of upper airway obstruction in children and adolescents.[1] One of the most concerning consequences of AH is its potential association with obstructive sleep apnea (OSA), a condition characterized by intermittent cessation or reduction of airflow during sleep.[2] If left untreated, OSA can lead to disrupted sleep, cognitive impairments, and various

Received: 3 September 2024
Accepted: 18 December 2024
Published: 27 December 2024

cardiovascular complications.[3,4] Therefore, accurate evaluation of AH is essential for the timely diagnosis and management of OSA in pediatric patients.

Currently, the most widely used methods for assessing AH are fiberoptic flexible nasopharyngoscopy and lateral cephalography, which rely on the Adenoid-Nasopharynx (A/N) ratio.[5,6] While these traditional approaches are effective, they have several limitations, including interobserver variability, dependence on clinician expertise, and sensitivity to factors such as patient positioning and image quality. These limitations can result in inconsistent measurements, underscoring the need for more objective, reproducible, and automated alternatives.

Deep learning (DL), a subset of machine learning (ML), employs artificial neural networks to learn and analyze the input data automatically, mimicking human cognitive processes.[7,8] This allows these networks to identify patterns and make predictions with remarkable accuracy. ML, as a broader domain, encompasses various algorithms that enable systems to learn from data without explicit programming.[9] Among these, DL has consistently demonstrated superior performance in medical applications, excelling in tasks such as image classification, segmentation, and diagnostic decision-making.[10] Notably, DL techniques have shown great promise in detecting and quantifying laryngeal conditions, including carcinoma, cancer, and vocal fold leukoplakia.[11–13] These achievements suggest that DL could also play a pivotal role in evaluating AH and its associated conditions.

Furthermore, DL has also been successfully applied to the analysis of sleep-related disorders, which are closely associated to AH. For instance, Abbasi et al utilized a multilayer perceptron neural network for EEG-based neonatal sleep-wake classification.[14] Similarly, a CNN-based decision support system was developed for detecting neonatal quiet sleep.[15] These studies demonstrate DL's ability to accurately analyze sleep patterns, highlighting its potential for enhancing sleep monitoring and early-stage development assessments. The increasing use of DL in these areas provides a strong foundation for applying its capabilities to improve the assessment of AH.

Several studies have explored the potential of DL in the evaluation of AH. For example, Shen et al validated a DL-based approach for classifying AH using lateral cephalograms, achieving promising results.[16] Bi et al introduced MIB-ANet, a multi-scale grading network, which showed robust performance in grading AH based on nasal endoscopy images.[17] While these advances mark significant process, challenges remain. Although DL algorithms have improved the accuracy of AH assessments, they often still rely on subjective input from clinicians, particularly in terms of training data labeling and model interpretation. This challenge is compounded by the inherent limitations of traditional evaluation methods. Furthermore, many DL models are trained using single neural networks, which can be prone to overfitting and may have reduced generalization to new data. The scarcity of large, high-quality datasets in specialized medical fields further hamper the clinical adoption of DL algorithms. These challenges underscore the need for innovative strategies to enhance the robustness, generalizability, and scalability of DL-based methods in AH assessment.

To address these issues, techniques such as transfer learning and ensemble learning have emerged as effective solutions. Transfer learning allows a model trained on one task to be adapted for a related task with a small dataset, leveraging knowledge gained from large-scale pre-trained models.[18] This approach has been shown to enhance model performance when data is scarce.[19] Ensemble learning, which combines multiple models to improve overall performance, has been used to aggregate predictions, thereby increasing accuracy and reduced overfitting.[20] Both techniques have been successfully applied in medical research to improve model robustness, even with limited training data.

In this study, (i) we propose a novel deep learning-based approach that integrates transfer learning and ensemble learning techniques to quantitatively assess AH in pediatric patients. (ii) Specifically, we develop a model that leverages transfer learning for accurate image segmentation and feature extraction from nasopharyngoscopy images. (iii) Our model is designed to automatically identify and quantify the A/N ratio, providing a more objective, reliable, and reproducible assessment of AH severity. (iv) We investigate the relationship between A/N ratio and AHI across different tonsil grades and explored the interaction among these variables. By improving the accuracy of these measurements, our approach aims to improve the diagnosis and management of AH and OSA, offering significant potential for clinical application.

Materials and Methods

Nasopharyngoscopy Images Collection

A total 1642 fiberoptic flexible nasopharyngoscopy images were collected from patients aged 3 to 12 years at the Department of Otolaryngology-Head and Neck Surgery, Zhongnan Hospital of Wuhan University. To ensure the quality of the data, images with obscured secretions or incomplete exposure of adenoids were systematically excluded from the dataset. Additionally, we excluded patients with severe nasopharyngeal anatomical abnormalities. Following this rigorous screening process, a total of 1500 images were retained for analysis. The images were acquired using a fiberoptic flexible nasopharyngoscopy (ENF-V3; OLYMPUS; Japan) and endoscopic video systems (OTV-S190; OLYMPUS; Japan), ensuring standardized imaging at a resolution of 516×531 pixels. This study was conducted in accordance with the Declaration of Helsinki and was approved by the Medical Ethics Committee of Zhongnan Hospital, Wuhan University (Approval No. 2022179K).

Evaluation of A/N Ratio by Otolaryngologists

Initially, two highly experienced otolaryngologists, each with 20 years of clinical experience, were engaged to evaluate the A/N ratio and grade AH using the collected nasopharyngoscopy images. To mitigate potential biases between the evaluations of two otolaryngologists, the Mann–Whitney *U*-test was conducted on their respective assessment outcomes. This statistical measure was used to ensure the reliability and objectivity of the evaluation process, providing a robust measure of consistency between the two evaluators.

Calculation of A/N Ratio in MATLAB

To improve the precision AH evaluations and ensure precise A/N ratio measurements, nasopharyngoscopy images were manually annotated using MATLAB's imfreehand tool. Two experienced otolaryngologists delineated the boundaries of the adenoid and the nasopharyngeal space to create region-specific masks. These annotations were then converted into binary masks for subsequent analysis.

However, manual outlining introduces potential biases. Variability can stem from inter-annotator inconsistencies over time, potentially leading to discrepancies in the final A/N ratio. To minimize these biases, all annotations were performed by trained otolaryngologists in image segmentation. Additionally, each image was independently annotated by two otolaryngologists. The final analysis relied on the overlapping region of their annotations. If the difference between the two otolaryngologists exceeded 5%, a senior otolaryngologist with over 30 years of clinical experience was consulted to perform re-annotation. This iterative process aimed to enhance the consistency and reliability.

To ensure reproducibility and account for any residual bias in the manual annotation process, the Mann–Whitney *U*-test was performed to evaluate the agreement between the two annotators. This analysis reinforced the reliability of the measurements and subsequent classification outcomes.

As illustrated in Figure 1A, the annotation process involved manually outlining the adenoid and nasopharyngeal regions, followed by assigning unique labels to each area: the background was labeled as 0 (black: 0,0,0), the unobstructed portion of the nasopharynx as 1 (gray: 128,128,128), and the adenoid region as 2 (white: 255,255,255). Using these annotations, the algorithm automatically computed the A/N ratio and categorized AH into three levels based on this ratio: small (<50%), medium (50–75%), and large (>75%).[21,22]

Deep Learning Architecture

MMSegmentation (MMSeg) was chosen to construct the DL framework for pixel-level classification due to its modularity, flexibility, and established performance across a variety of segmentation tasks. It was publicly available on the GitHub repository.[23] MMSeg provides a unified benchmark and supports a broad range of state-of-the-art architectures that are widely recognized for their effectiveness in semantic segmentation. Specifically, the framework's modularity allows us to experiment with multiple components of the segmentation pipeline, including backbones, heads, and loss functions, ensuring that we could easily tailor the architecture to the specific needs of our dataset and task (Figure 1B).

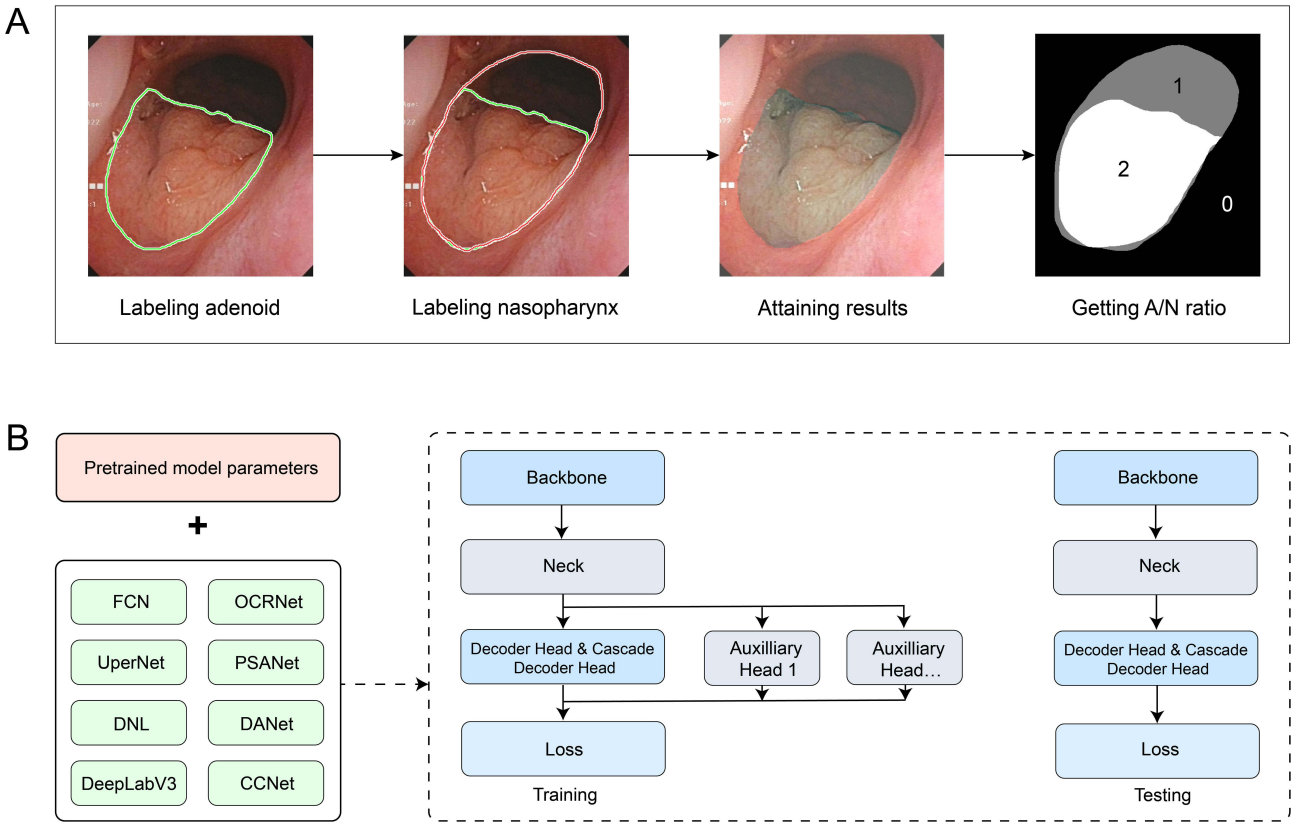

Figure 1 Calculation process of the A/N ratio using MATLAB and deep learning model architecture. (**A**) A fiberoptic nasopharyngoscopy image was selected, highlighting the adenoid cross-sectional area and the nasopharynx. The developed algorithm was then applied to automatically calculate the A/N ratio. (**B**) The network architecture used for both training and testing stages consisted of identical components, including essential modules such as Backbone, Neck, Decoder Head, and Loss, along with optional modules like the Neck and Auxiliary Head.

MMSeg supports several popular and contemporary frameworks, such as ResNeSt,[24] UNet,[25] DeepLabV3,[26] FastFCN.[27] The selected architectures were chosen not only because of their proven success in similar tasks but also because they provide different strengths in feature extraction, spatial information retention, and multi-scale processing. By combining these components, different models and a customized semantic segmentation framework could be easily constructed.

Transfer Learning

Given the limited size of medical datasets, we leveraged transfer learning to overcome the challenge of training with scarce annotated data. Pre-trained models from MMSeg, originally trained on large-scale semantic segmentation datasets like Cityscapes and ADE20K, provided a strong starting point for our training. These models already possess rich, high-level feature representations that we could fine-tune for our specific medical segmentation task. By adjusting the final layers to match our dataset's categories, we ensured that the models were optimized for our classification needs while significantly reducing training time and improving accuracy, especially when compared to training models from scratch.

Ensemble Learning

To further enhance the robustness and accuracy of our segmentation models, we employed ensemble learning. Ensemble learning is a technique that combines multiple learning algorithms to achieve superior predictive performance compared to any individual algorithm.[28] In this study, we employed ensemble learning using eight distinct network architectures to develop the final network, named SUMNet. This approach improved the accuracy and stability of the model while mitigating the risks of overfitting and underfitting.

Model Selection and Training Details

In this study, DL experiments were conducted using a curated set of frameworks from the MMSeg Library, selecting frameworks based on their reported Mean Intersection over Union (MIoU) metrics on the Cityscapes Datasets.[29] Consequently, eight networks were chosen: FCN,[30] UPerNet,[31] DNLNet,[32] DeepLabV3,[26] OCRNet,[33] PSANet,[34] DANet[35] and CCNet,[36] ranked in decreasing order of their best MIoU scores.

To ensure an unbiased evaluation of model performance, the images were randomly partitioned into a training set (4/6), a validation set (1/6), and a test set (1/6). Various data augmentation techniques, including Random Resized Crop, Random Flip, Normalize, Photometric Distortion, and Padding, were employed during training to increase the diversity of training data and improve the model's generalization ability. For testing, resize, random flip, and normalization were applied. Additionally, an optimizer with a poly learning strategy was employed, which effectively minimized the loss function during training. The loss function used was cross-entropy loss, commonly employed in classification tasks. Each network was trained for a total of 500 epochs to ensure sufficient convergence, and performance was evaluated on the test set.

Performance Assessment

Precision, Recall and Intersection Over Union

In our multi-class, pixel-level classification task, precision and recall were the primary metrics for evaluating model performance. Precision is defined as the proportion of true positives among all instances predicted as positive, while recall measures the proportion of true positives among all instances that actually belong to the class. Intersection over Union (IoU) was used to quantify the overlap between the predicted and ground-truth segmentation at each pixel. The formulas are as follows:

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$IoU = \frac{TP}{(TP + FP + FN)}$$

Where TP represents true positives, FP represents false positives, and FN represents false negatives. The image was segmented into three categories, represented by pixels 0, 1, and 2. These metrics were calculated by performing a pixel-by-pixel comparison between the predicted and reference images in MATLAB. Additionally, the average precision (AP), average recall (AR), overall accuracy (OA), and MIoU were computed to provide a comprehensive evaluation of the DL model's performance.

Cohen's Kappa

Cohen's kappa coefficient is a statistical measure used to assess the consistency between evaluators in classification tasks. In DL applications, it evaluates the agreement between model predictions and true labels, providing insight into the model's accuracy and stability while mitigating bias from class imbalance. A higher kappa coefficient indicates greater consistency between the model's predictions and the true labels, reflecting improved performance. In this study, Cohen's kappa was used to assess the effectiveness of the DL framework. P_O is the observed agreement, and Pe is the expected agreement. The formula is as follows:

$$\kappa = \frac{\mathrm{Po} - \mathrm{Pe}}{1 - \mathrm{Pe}}$$

Confusion Matrix

The confusion matrix is a widely used tool for evaluating the performance of supervised learning algorithms in classification tasks. In this study, each labeled or predicted image was assigned a grade label based on its computed

A/N ratio. A confusion matrix was then constructed by comparing the assigned grade labels of labeled images with those of their corresponding predicted images.

Receiver Operating Characteristic Curve

To evaluate the performance of the DL method in assessing AH compared to human experts, the area under the curve (AUC) of the Receiver operating characteristic (ROC) curve was utilized as a key metric. The AUC, which ranging from 0 to 1, reflects the accuracy of classification, with values closer to 1 indicating superior performance and a higher reliability for the DL model.

Polysomnography

Patient information obtained from nasopharyngoscopy images was used to collect corresponding polysomnography (PSG) data for the included subjects. Due to the absence of PSG data for some patients following after nasopharyngoscopy examination, valid data were ultimately obtained for 1000 out of an initial 1500 participants. Polysomnographic recordings were conducted using the Philips Alice 6 multi-channel sleep system, which continuously and synchronously monitored various physiological parameters throughout a full night of sleep. These parameters included electrocardiogram (ECG), electroencephalogram (EEG), electromyography (EMG), oral and nasal airflow, chest and abdominal movements, body position, and finger pulse oximetry.

Sleep staging and respiratory event analysis were performed in accordance with the guidelines of the American Academy of Sleep Medicine (AASM). The recordings were carefully reviewed and interpreted by trained technicians, who extracted key indicators such as the apnea-hypopnea index (AHI). The AHI was calculated as the number of apneas and hypopneas per hour of sleep.

To evaluate the clinical relevance of the A/N ratio, as determined by the DL method, in diagnosing OSA, a Spearman correlation analysis was performed between the A/N ratio and the corresponding AHI values. To address the potential confounding effect of tonsil hypertrophy on AHI measurements, tonsil size information was collected from all 1000 participants. Tonsil size was graded on a standardized scale as follow: 0 (in Fossa), +1 (<25%), +2 (25%~50%), +3 (50%~75%) and +4 (>75%).[37]

Results

Evaluation by Otolaryngologists and MATLAB

The reliability of subjective assessments of the A/N by otolaryngologists and the MATLAB-based calibration was analyzed using the Mann–Whitney *U*-test in GraphPad Prism 9. A significance threshold of $p>0.05$ was considered to indicate no significant difference. The subjective evaluations performed by otolaryngologists resulted in a *p*-value of 0.8507, suggesting no significant differences between the two otolaryngologists, thus supporting the reliability of subjective assessment as a valid reference for manual evaluation. Additionally, MATLAB calibration results demonstrated a *p* value of 0.679, indicating no significant differences between their outcomes. These findings confirm the robustness and reliability of the MATLAB algorithm in calculating A/N ratio (Figure 2).

Deep Learning Model Performance

Eight DL frameworks were employed to train a segmented training set, which was subsequently tested on a separate test set. Additionally, the ensemble method SUMNet was evaluated for its performance. As shown in Table 1 and Figure 3, the models were assessed using multiple metrics, including precision, recall, AP, AR, MIoU, OA, and Cohen's kappa. Among the individual models, FCN, UperNet, and DeepLabV3 consistently demonstrated high performance, with precision values ranging from 0.9503 to 0.9558 and kappa values between 0.62 and 0.82. Overall, all eight models performed well across the evaluated metrics, confirming the effectiveness of DL methods in assessing AH. Notably, SUMNet, an ensemble learning model, achieved slight improvements in several metrics. It attained the highest precision (0.9616), recall (0.7538), AP (0.9153), AR (0.9290), MIoU (0.8519), OA (0.9243), and kappa (0.87). These findings indicate that the ensemble approach provided modest enhancement in model performance, particularly in terms of

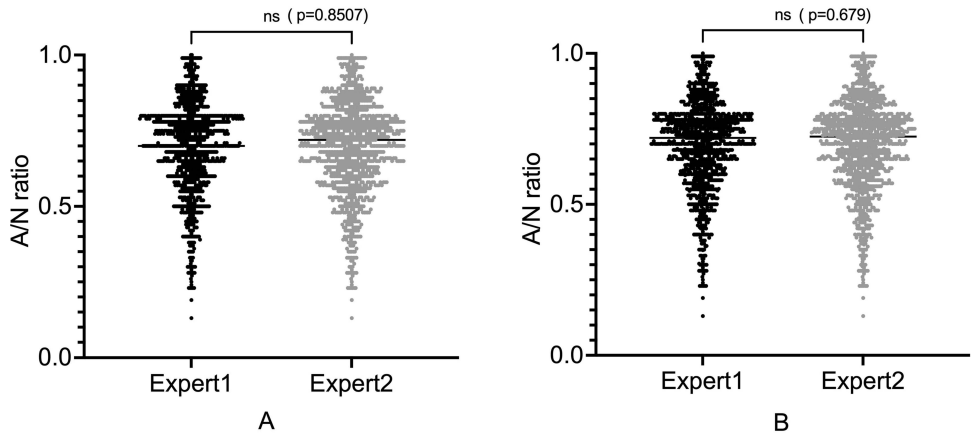

Figure 2 Scatterplot of the Mann–Whitney *U*-test for A/N ratio evaluations. (**A**) Each scatter point represents an individual A/N ratio value assessed by the two experts. (**B**) Each scatter point corresponds to an individual A/N ratio value calculated by the MATLAB algorithm, based on calibration performed by the same two experts. This visual representation offers a comprehensive comparison between the expert evaluations and the MATLAB-calculated A/N ratio values.

consistency and overall accuracy. SUMNet's robust performance underscores its potential as an effective and objective tool for quantitative evaluation of AH.

Classification Results of the Deep Learning-Based MMseg

Following the established methodology for evaluating AH degrees,[22] confusion matrices were generated for the predictions made by the eight individual networks and ensemble network. These confusion matrices provided a comprehensive evaluation of classification accuracy and discrimination ability across three AH categories (Figure 4). All frameworks demonstrated strong performances, validating the effectiveness of DL methods in grading AH, particularly in the medium and large categories. Among the models, the ensemble learning method, SUMNet, achieved the most balanced performance across all categories, correctly predicting 13 small, 98 medium, and 93 large adenoids. Notably, SUMNet exhibited reduced confusion between medium and large categories compared to the individual networks, indicating superior consistency in classification.

ROC curve analysis was performed to compare the performance of DL models against expert evaluations, with the AUC serving as an indicator of overall classification ability (Figure 5). The DL models consistently outperformed expert evaluations, achieving AUC values ranging from 0.83 to 0.96, compared to an AUC of 0.74 for expert assessments. SUMNet achieved an AUC of 0.85, further underscoring its robust classification capabilities. These findings underscore

Table 1 The Performance Metrics of Deep Learning Frameworks for a/N Ratio Assessment

Network	Precision			Recall			AP	AR
	0	1	2	0	1	2		
FCN	0.9535	0.7648	0.9103	0.9301	0.8436	0.9126	0.8762	0.8954
UperNet	0.9521	0.7494	0.9157	0.9285	0.8512	0.9084	0.8724	0.8960
DNLNet	0.9545	0.7568	0.9114	0.9317	0.8156	0.9207	0.8742	0.8893
DeepLabV3	0.9544	0.7575	0.9087	0.9303	0.8096	0.9231	0.8735	0.8877
OCRNet	0.9558	0.7421	0.9112	0.9249	0.8451	0.9149	0.8697	0.8950
PSANet	0.9518	0.7552	0.9111	0.9303	0.8194	0.9159	0.8727	0.8885
DANet	0.9503	0.7588	0.9095	0.9316	0.8119	0.9146	0.8728	0.8860
CCNet	0.9481	0.7551	0.9125	0.9321	0.8082	0.9133	0.8719	0.8845
SUMNet	0.9616	0.7538	0.9153	0.9290	0.8519	0.9243	0.8769	0.9017

Notes: 0: background; 1: the unblocked part of nasopharynx; 2: adenoid.
Abbreviations: AP, average precision; AR, average recall.

https://doi.org/10.2147/NSS.S492146

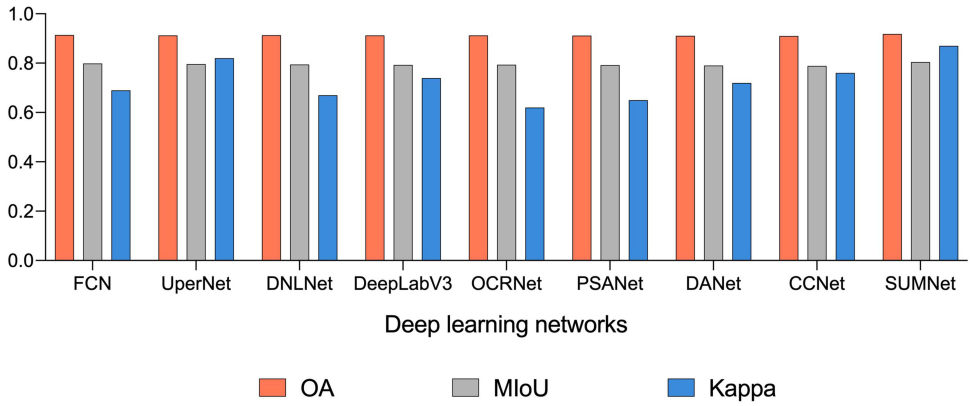

Figure 3 Performance of different deep learning networks. SUMNet, an ensemble learning model, demonstrated slight improvements in several metrics, including overall accuracy (OA), mean intersection over union (MIoU), and Cohen's kappa (Kappa).

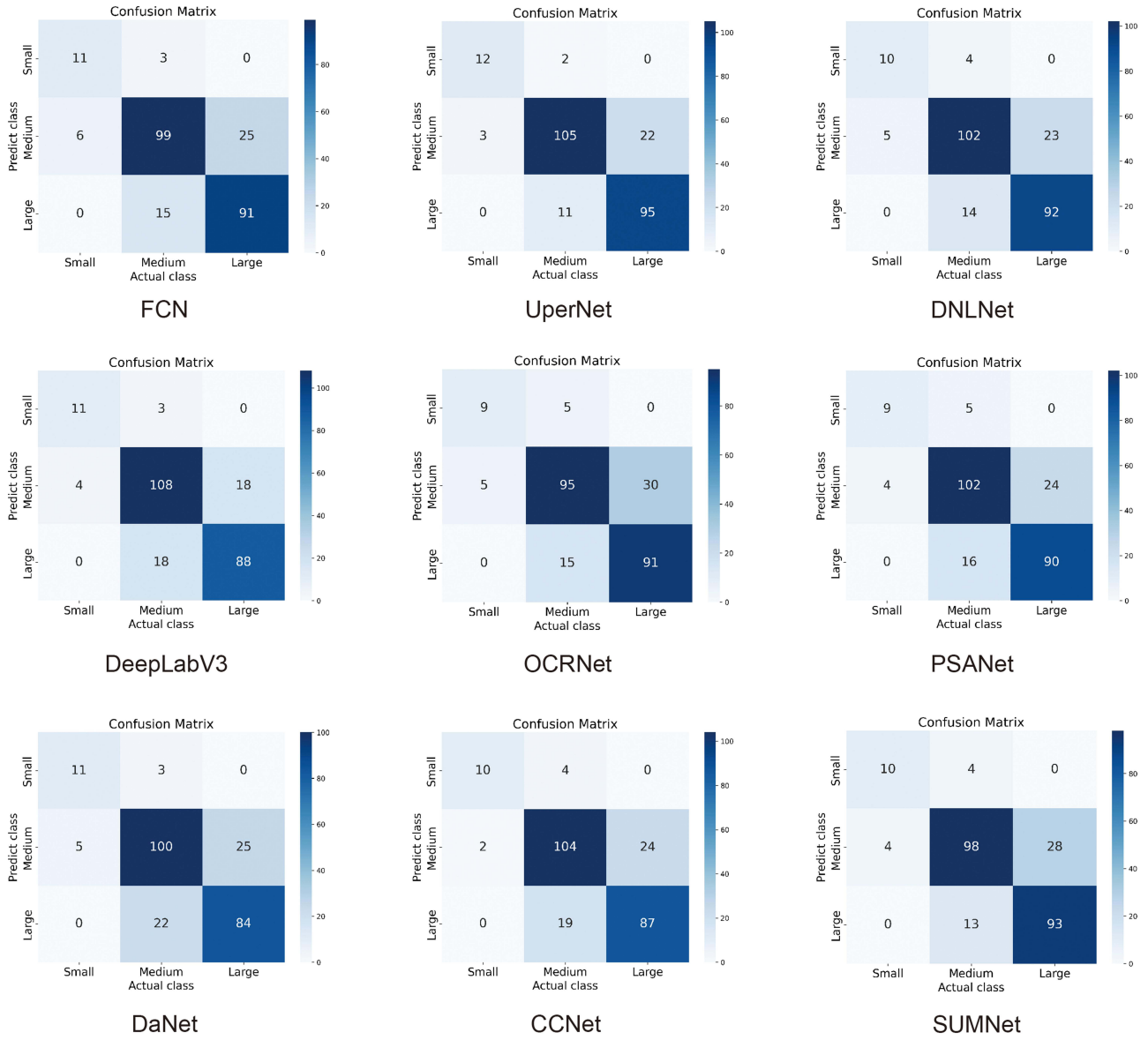

Figure 4 Confusion matrix of adenoid hypertrophy degree performance for deep learning method and MATLAB algorithm. Adenoid hypertrophy degree is classified into three categories: small (A/N ratio 0–50%), medium (A/N ratio 50–75%), and large (A/N ratio 75–100%). In each confusion matrix, the horizontal axis represents the MATLAB results (actual class), while the vertical axis represents the deep learning results (predicted class).

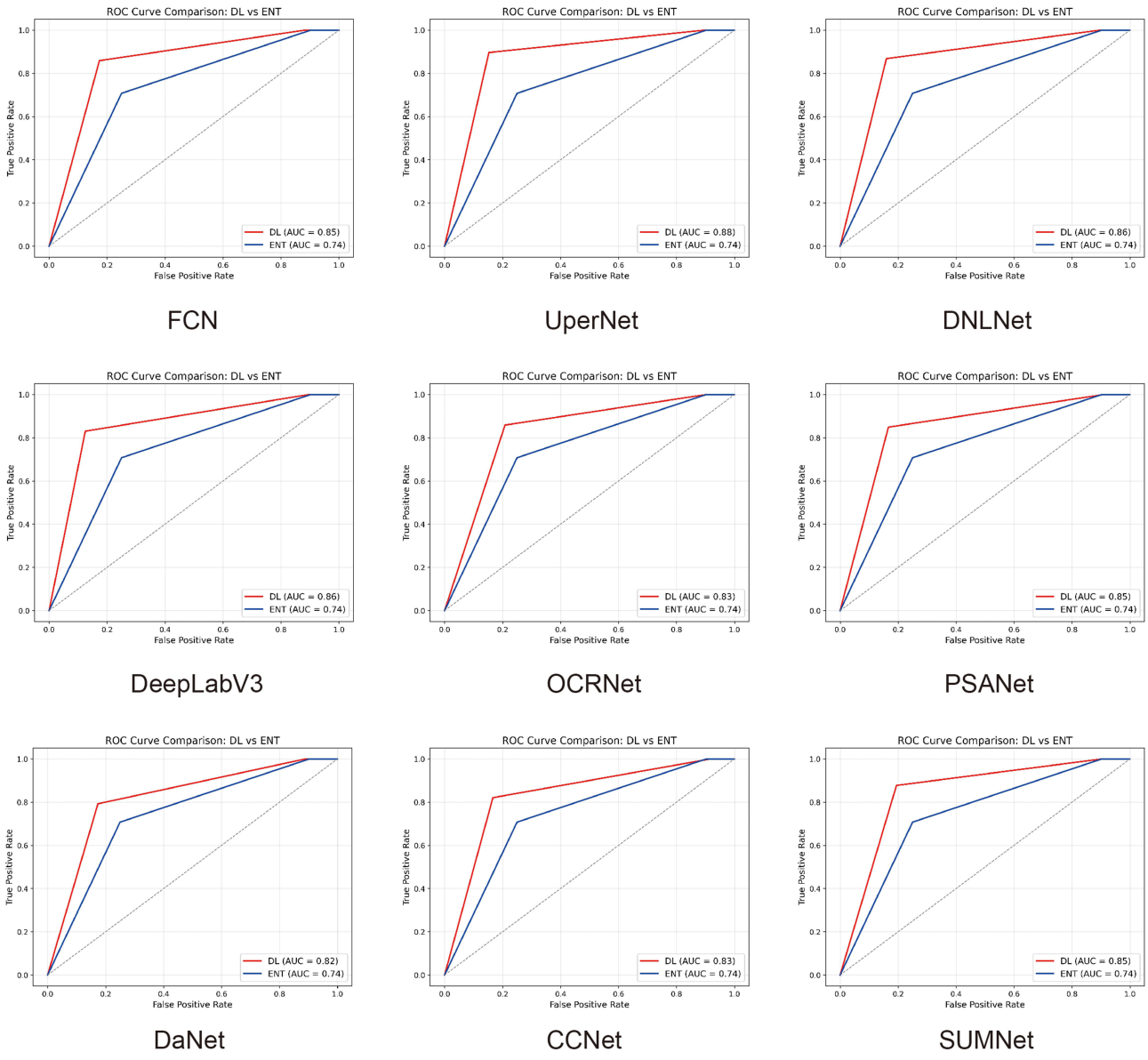

Figure 5 ROC curves of different deep learning models and human experts. Adenoid hypertrophy degree is classified into three categories: small (A/N ratio 0–50%), medium (A/N ratio 50–75%), and large (A/N ratio 75–100%). The results show the AUC values for both deep learning models and human experts. Deep learning models consistently outperformed expert evaluations in terms of performance.

the effectiveness of DL models, particularly SUMNet, in accurately assessing AH. The results demonstrate the superiority of SUMNet over individual models and expert evaluations, making it a reliable tool for the objective classification of AH degrees.

Correlation Between A/N Ratio and AHI Across Tonsillar Sizes

To investigate the relationship between the A/N ratio and AHI, we utilized both the A/N ratio subjectively assessed by experts and that calculated by the DL-based SUMNet model. Spearman correlation analyses were conducted to evaluate the association between the A/N ratio and AHI across different tonsillar sizes (Figure 6).

Scatter plots depicted the correlations for pediatric patients with different tonsillar grades. No cases with tonsillar grade 0 (in fossa) were identified; thus, data for this category were not included. In patients with tonsillar grade +1 (<25%), a strong positive correlation was observed for both SUMNet-derived ratios (r=0.9052) and expert assessments (r=0.4590), indicating a significant increase in AHI with higher A/N ratios. For grade +2 tonsils (25–50%), the

https://doi.org/10.2147/NSS.S492146

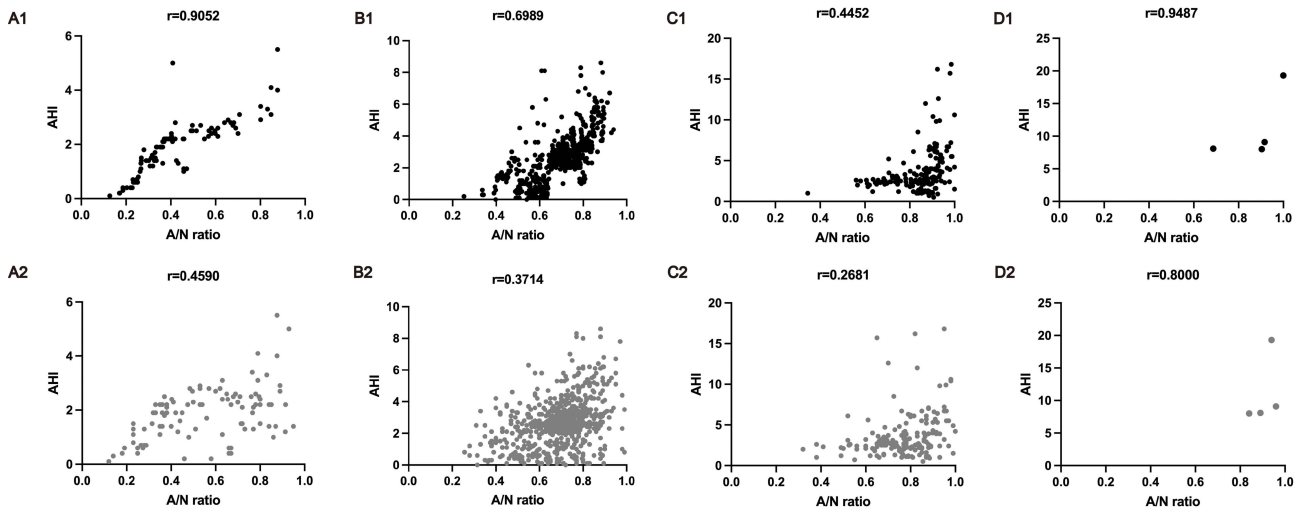

Figure 6 Scatter plot of Spearman correlation coefficient between the A/N ratio and AHI across different tonsillar grades. (**A1**–**D1**) Correlation between A/N ratio and AHI calculated by the deep learning-based SUMNet model. (**A2**–**D2**) Correlation between A/N ratio and AHI calculated by human experts. The Spearman correlation coefficient, ranging from −1 to 1, indicates the strength and direction of the relationship. A coefficient close to 1 reflects a strong positive correlation, close to −1 indicates a strong negative correlation, and close to 0 suggests no linear correlation. (**A1** and **A2**): In patients with tonsillar grade +1 (<25%). (**B1** and **B2**) In patients with tonsillar grade +2 (25–50%). (**C1** and **C2**) In patients with tonsillar grade +3 (50%-75%). (**D1** and **D2**) In patients with tonsillar grade +4 (>75%).

correlation weakened to moderate level (r=0.6989 for SUMNet and r=0.3714 for expert evaluations). Similarly, for grade +3 tonsils (50–75%), a weak positive correlation was observed (r=0.4452 for SUMNet and r=0.2681 for expert evaluations), suggesting a diminishing influence of the A/N ratio on AHI as tonsillar size increases. Interestingly, for patients with grade +4 tonsils (>75%), a very strong positive correlation was observed (r=0.9487 for SUMNet and r=0.8000 for expert evaluations), emphasizing the substantial influence of both adenoid and tonsillar hypertrophy on AHI in cases of severe enlargement.

Overall, the A/N ratio, whether calculated by the DL model or determined through expert assessments, demonstrated a robust correlation with AHI, underscoring its potential as a valuable reference for the pre-diagnosis of OSA. The superior performance of the DL model in producing higher correlation coefficients further highlights its reliability. Moreover, the gradual decline in correlation coefficients for tonsillar grades +1 to +3 suggests that tonsillar hypertrophy significantly influences AHI, particularly in moderate cases.

Discussion

This study developed and validated a high-performance DL algorithm for the quantitative assessment of AH and explored its correlation with AHI across different tonsillar sizes. The findings demonstrate the efficacy of DL frameworks in providing accurate and reliable assessments, as well as highlighting the relationship between AH, AHI, and tonsillar size.

The results presented in Table 1 clearly demonstrate that all eight DL frameworks exhibit strong performance across various evaluation metrics. Previous studies have also utilized DL methods to assess AH and its grading using nasopharyngoscopy images. For example, Bi et al achieved an F1 score of 0.7625 and an accuracy of 0.7680 using the MIB-ANet model.[17] Similarly, Zheng et al evaluated four classic convolutional neural network (CNN) architectures —AlexNet, VGG16, ResNet50, and GoogleNet—alongside MIB-ANet, and found that MIB-ANet outperformed the other models.[38] In comparison, the DL frameworks assessed in this study achieved accuracy values exceeding 0.9, with AP and AR values above 0.86. These results highlight the superior performance of our approach. The accuracy of the A/N ratio is critical for grading AH and directly influences clinical decision-making, particularly in the diagnosis and treatment planning for children with OSA. The enhanced accuracy of our DL models offered significant potential to support clinicians in making more precise and reliable evaluations of AH.

The adenoid are adjacent to Eustachian tube, the posterior nasal septum, and soft palate, and are often influenced by surrounding tissue structures and secretions during clinical evaluation.[39] To assess model accuracy in image segmentation, particularly in predicting and classifying region boundaries, the MIoU metric was employed. Most MIoU values were around 0.8, indicating effective boundary delineation. Additionally, the Kappa coefficient was calculated to evaluate the agreement between model predictions and ground truth values. SUMNet achieved a Kappa value of 0.87, demonstrating strong consistency with the ground truth. Notably, despite using a relatively small dataset, our approach outperformed several studies that employed larger datasets. This superior performance was attributed to the adoption of the MMSeg framework, which incorporates transfer learning techniques. By leveraging pre-trained models, this framework enhanced the model's generalization capability, thereby improving segmentation accuracy. Furthermore, the MMSeg framework enables more precise delineation of the adenoid in nasopharyngoscopy images, reducing segmentation errors.

Figure 4 revealed that all DL models performed effectively in classifying adenoid sizes, particularly in the "medium" category. This finding suggests that the DL models developed in this study can accurately and directly assess the grade of AH. Additionally, ROC curve analysis demonstrated that the AUC values for all DL models surpassed those achieved by expert evaluations. This result highlights the superior performance of the DL models compared to subjective clinician assessments in determining the degree of AH. According to the guidelines of the American Academy of Pediatrics, adenotonsillectomy is recommended as the first-line treatment for children diagnosed with OSA if clinical examination confirms adenotonsillar hypertrophy and no surgical contraindications exist. Moreover, the severity of AH plays a critical role in determining whether PSG is required as the next diagnostic step. The DL models developed in this study provide substantial value in clinical practice by enabling reliable and objective assessments of AH. These models offer robust support for clinical decision-making and facilitate the development of individualized diagnostic and treatment plans.

It is important to note that individual models may produce variable results due to inherent idiosyncrasies in the training data and the stochastic nature of the model initialization. This variability underscores the potential advantages of ensemble learning techniques in improving model stability and robustness.[20] In our investigation, the ensemble learning method, SUMNet, demonstrated superior performance compared to all other models, achieving precision of 0.9616, an AP of 0.9153, an MIoU of 0.8519, and OA of 0.9182. Although the observed improvements were modest, this outcome is likely attributable to the already strong performance of the individual models. Consequently, the additional benefits provided by ensemble method were limited, as the base models used in SUMNet were already highly effective. While ensemble methods can enhance the stability and robustness of predictions, its ability to further boost performance is constrained when the base models have already achieved high accuracy. The relatively small incremental improvements observed in this study suggest that the ensemble method primarily contributed complementary information, rather than generating a substantial performance enhancement, due to the optimized nature of the base models.

To objectively assess the relationship between the A/N ratio and AHI while minimizing confounding effects from tonsil hypertrophy, we analyzed these metrics across different tonsil grades. A positive correlation between the A/N ratio and AHI was observed across all tonsil size grades, regardless of whether the ratio was derived from DL models or expert evaluations. However, as tonsil grades increased from +1 to +3, the correlation coefficient gradually decreased, suggesting that tonsil hypertrophy becomes an increasingly significant factor influencing AHI. Despite the limited sample size for grade +4 tonsils, the consistently high AHI values observed underscore the critical role of severe tonsil hypertrophy in pediatric OSA. Notably, the DL model consistently yielded higher correlation coefficients than expert evaluations, further demonstrating its superiority. These findings highlight the importance of objective evaluations of AH, such as those provided by DL models, in comprehensively assessing OSA in children.

Several factors contributed to the strong performance of the DL approach in this study. First, the small dataset size made transfer learning an effective strategy, as pre-trained networks minimized reliance on target domain data and reduced computational demands.[18] Second, even advanced DL networks often yield variable results due to differences in training strategies and data properties.[23,40] Third, many medical DL models rely on modular combinations of existing architectures rather than novel innovations, which may not always outperform established models. Given that this study aimed to compute the A/N ratio for clinical use, minor deviations in accuracy were deemed acceptable. Therefore, we

prioritized well-optimized networks, which provided robust performance while maintaining generalizability for medical image semantic segmentation tasks with limited datasets.

Limitations and Future Works

This study has several limitations. First, the generalizability of the findings to larger datasets and real-world clinical settings remains uncertain, primarily due to the relatively small dataset used for model training. This limitation raises concerns about potential overfitting, where the model may perform well on the training set but fail to generalize effectively to other imaging conditions. Future research should validate these frameworks on larger, more diverse datasets to ensure their robustness across varying demographics and clinical environments. Second, inconsistent patient adherence to post-operative PSG examinations limited our ability to comprehensively evaluate PSG improvements following adenotonsillectomy. This gap underscores the need for more follow-up protocols in future studies to assess long-term treatment outcomes. Third, this study did not account for other contributing factors to pediatric OSA, such as obesity, craniofacial anomalies, and nasal obstruction. Including these variables in future research could provide a more comprehensive understanding the multifactorial nature of pediatric OSA and further refine the clinical utility of DL models. Addressing these limitations will be crucial for translating these models into practical, real-time clinical applications.

Conclusions

In conclusion, this study developed a novel DL algorithm that integrates transfer learning and ensemble learning to enhance the accuracy of AH quantification from fiberoptic nasopharyngoscopy images. Additionally, we investigated the association between AH and AHI, providing valuable insights into the relationship between AH and the severity of OSA. Our findings offer an objective, reliable method for evaluating AH and highlight its potential as a clinically significant marker for OSA diagnosis. These advancements underscore the utility of leveraging DL in refining diagnostic approaches for pediatric patients with OSA.

Ethics Declaration

This study was conducted in accordance with the Declaration of Helsinki and was approved by the Medical Ethics Committee of Zhongnan Hospital, Wuhan University (Approval No. 2022179K). Informed consent was obtained from the legal guardians of all study participants.

Data Sharing Statement

The fiberoptic nasopharyngoscopy images data in this study are not publicly available for patient privacy purposes but are available from the corresponding authors upon reasonable request. The source codes are provided at GitHub https://github.com/open-mmlab/mmsegmentation/.

Funding

This study was supported by National Natural Science Foundation of China (82071033).

Disclosure

All authors included in this research declare no competing interests.

References

1. Pereira L, Monyror J, Almeida FT, et al. Prevalence of adenoid hypertrophy: a systematic review and meta-analysis. *Sleep Med Rev*. 2018;38:101–112. doi:10.1016/j.smrv.2017.06.001
2. Zhu Y, Wang S, Yang Y, et al. Adenoid lymphocyte heterogeneity in pediatric adenoid hypertrophy and obstructive sleep apnea. *Front Immunol*. 2023;14:1186258. doi:10.3389/fimmu.2023.1186258
3. Lévy P, Kohler M, McNicholas WT, et al. Obstructive sleep apnoea syndrome. *Nat Rev Dis Primers*. 2015;1:15015. doi:10.1038/nrdp.2015.15
4. Arnaud C, Bochaton T, Pépin JL, Belaidi E. Obstructive sleep apnoea and cardiovascular consequences: pathophysiological mechanisms. *Arch Cardiovasc Dis*. 2020;113(5):350–358. doi:10.1016/j.acvd.2020.01.003

5. Ahmad Z, Krüger K, Lautermann J, et al. Adenoid hypertrophy-diagnosis and treatment: the new S2k guideline. *Hno*. 2023;71(Suppl 1):67–72. doi:10.1007/s00106-023-01299-6
6. Baldassari CM, Choi S. Assessing adenoid hypertrophy in children: x-ray or nasal endoscopy? *Laryngoscope*. 2014;124(7):1509–1510. doi:10.1002/lary.24366
7. Dong S, Wang P, Abbas K. A survey on deep learning and its applications. *Comput Sci Rev*. 2021;40:100379. doi:10.1016/j.cosrev.2021.100379
8. Janiesch C, Zschech P, Heinrich K. Machine learning and deep learning. *Electron Mark*. 2021;31(3):685–695. doi:10.1007/s12525-021-00475-2
9. Sarker IH. Machine learning: algorithms, real-world applications and research directions. *SN Comput Sci*. 2021;2(3):160. doi:10.1007/s42979-021-00592-x
10. Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. *Med Image Anal*. 2017;42:60–88. doi:10.1016/j.media.2017.07.005
11. Wang Z, Fang M, Zhang J, et al. Radiomics and deep learning in nasopharyngeal carcinoma: a review. *IEEE Rev Biomed Eng*. 2024;17:118–135. doi:10.1109/rbme.2023.3269776
12. Xu ZH, Fan DG, Huang JQ, Wang JW, Wang Y, Li YZ. Computer-aided diagnosis of laryngeal cancer based on deep learning with laryngoscopic images. *Diagnostics*. 2023;13(24):3669. doi:10.3390/diagnostics13243669
13. 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(12):3129–3145. doi:10.1002/hed.27543
14. Abbasi SF, Ahmad J, Tahir A, et al. EEG-based neonatal sleep-wake classification using multilayer perceptron neural network. *IEEE Access*. 2020;8:183025–183034. doi:10.1109/ACCESS.2020.3028182
15. Abbasi SF, Abbasi QH, Saeed F, Alghamdi NS. A convolutional neural network-based decision support system for neonatal quiet sleep detection. *Math Biosci Eng*. 2023;20(9):17018–17036. doi:10.3934/mbe.2023759
16. Shen Y, Li X, Liang X, et al. A deep-learning-based approach for adenoid hypertrophy diagnosis. *Med Phys*. 2020;47(5):2171–2181. doi:10.1002/mp.14063
17. Bi M, Zheng S, Li X, et al. MIB-ANet: a novel multi-scale deep network for nasal endoscopy-based adenoid hypertrophy grading. *Front Med*. 2023;10:1142261. doi:10.3389/fmed.2023.1142261
18. Weiss K, Khoshgoftaar TM, Wang D. A survey of transfer learning. *J Big Data*. 2016;3(1):1–40. doi:10.1186/s40537-016-0043-6
19. Zhuang F, Qi Z, Duan K, et al. A comprehensive survey on transfer learning. *Proc IEEE*. 2021;109(1):43–76. doi:10.1109/JPROC.2020.3004555
20. Dong X, Yu Z, Cao W, Shi Y, Ma Q. A survey on ensemble learning. *Front Comput Sci*. 2020;14(2):241–258. doi:10.1007/s11704-019-8208-z
21. Wormald PJ, Prescott CA. Adenoids: comparison of radiological assessment methods with clinical and endoscopic findings. *J Laryngol Otol*. 1992;106(4):342–344. doi:10.1017/s0022215100119449
22. Elwany S. The adenoidal-nasopharyngeal ratio (AN ratio). Its validity in selecting children for adenoidectomy. *J Laryngol Otol*. 1987;101(6):569–573. doi:10.1017/s0022215100102269
23. Chen K, Wang J, Pang J, et al. MMSegmentation: openMMLab semantic segmentation toolbox and benchmark. *arXiv*; 2019.doi:10.48550/arXiv.1906.07155.
24. Zhang H, Wu C, Zhang Z, et al. ResNeSt: split-attention networks. *arXiv*. 2020. doi:10.48550/arXiv.2004.08955.
25. Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. *Lect Notes Comput Sci*. 2015;9351:234–241. doi:10.1007/978-3-319-24574-4_28
26. Chen LC, Papandreou G, Schroff F, Adam H. Rethinking atrous convolution for semantic image segmentation. *arXiv*. 2017. doi:10.48550/arXiv.1706.05587
27. Wu H, Zhang J, Huang K, Liang K, Yu Y. FastFCN: rethinking dilated convolution in the backbone for semantic segmentation. arXiv. 2019. doi:10.48550/arXiv.1903.11816
28. Rokach L. Ensemble-based classifiers. *Artif Intell Rev*. 2010;33(1):1–39. doi:10.1007/s10462-009-9124-7
29. Cordts M, Omran M, Ramos S, et al. The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).NV, USA: Las Vegas; 2016:3213–3223. doi:10.1109/CVPR.2016.350
30. Shelhamer E, Long J, Darrell T. Fully convolutional networks for semantic segmentation. *IEEE Trans Pattern Anal Mach Intell*. 2017;39(4):640–651. doi:10.1109/tpami.2016.2572683
31. Xiao T, Liu Y, Zhou B, Jiang Y, Sun J. *Unified Perceptual Parsing for Scene Understanding*. Cham: Springer International Publishing; 2018:418–434. doi:10.48550/arXiv.1807.10221
32. Yin M, Yao Z, Cao Y, et al. *Disentangled Non-Local Neural Networks*. Springer; 2020. doi:10.48550/arXiv.2006.06668
33. Yuan Y, Chen X, Wang J. Object-contextual representations for semantic segmentation. *Lect Notes Comput Sci*. 2020;173–190. doi:10.1007/978-3-030-58539-6_11
34. Zhao H, Zhang Y, Liu S, et al. Psanet: point-wise spatial attention network for scene parsing. *Lect Notes Comput Sci*. 2018:267–283. doi:10.1007/978-3-030-01240-3_17
35. Fu J, Liu J, Tian H, et al. Dual attention network for scene segmentation. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA; 2019:3141–3149. doi:10.1109/CVPR.2019.00326
36. Huang Z, Wang X, Wei Y, et al. CCNet: criss-cross attention for semantic segmentation. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South). 2019:603–612. doi:10.1109/ICCV.2019.00069
37. Brodsky L. Modern assessment of tonsils and adenoids. *Pediatr Clin North Am*. 1989;36(6):1551–1569. doi:10.1016/s0031-3955(16)36806-7
38. Zheng S, Li X, Bi M, et al. Contrastive learning-based adenoid hypertrophy grading network using nasoendoscopic image. In: 2022 IEEE 35th International Symposium on Computer-Based Medical Systems (CBMS). 2022:377–382. doi:10.1109/CBMS55023.2022.00074
39. Parikh SR, Coronel M, Lee JJ, Brown SM. Validation of a new grading system for endoscopic examination of adenoid hypertrophy. *Otolaryngol Head Neck Surg*. 2006;135(5):684–687. doi:10.1016/j.otohns.2006.05.003
40. Chouai M, Dolezel P, Stursa D, Nemec Z. New end-to-end strategy based on DeepLabv3+ semantic segmentation for human head detection. *Sensors*. 2021;21(17):5848. doi:10.3390/s21175848