Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

A fully-automatic semi-supervised deep learning model for difficult airway assessment

Guangzhi Wang [a,1], Chenxi Li [a,1], Fudong Tang [a], Yangyang Wang [a], Su Wu [a], Hui Zhi [a], Fan Zhang [b], Meiyun Wang [c], Jiaqiang Zhang [a,*]

[a] *Department of Anesthesiology and Perioperative Medicine, People's Hospital of Zhengzhou University, Henan Provincial People's Hospital, Zhengzhou, Henan, China*

[b] *Unmanned System Research Institute, Northwestern Polytechnical University, Xi'an, China*

[c] *Department of Medical Imaging, People's Hospital of Zhengzhou University, Henan Provincial People's Hospital, Zhengzhou, Henan, China*

ARTICLE INFO

Keywords:
Artificial intelligence
Deep learning
Difficult airway
Elective surgery
General anesthesia

ABSTRACT

Background: Difficult airway conditions represent a substantial challenge for clinicians. Predicting such conditions is essential for subsequent treatment planning, but the reported diagnostic accuracies are still quite low. To overcome these challenges, we developed a rapid, non-invasive, cost-effective, and highly-accurate deep-learning approach to identify difficult airway conditions through photographic image analysis.

Methods: For each of 1000 patients scheduled for elective surgery under general anesthesia, images were captured from 9 specific and different viewpoints. The collected image set was divided into training and testing subsets in the ratio of 8:2. We used a semi-supervised deep-learning method to train and test an AI model for difficult airway prediction.

Results: We trained our semi-supervised deep-learning model using only 30% of the labeled training samples (with the remaining 70% used without labels). We evaluated the model performance using metrics of accuracy, sensitivity, specificity, F1-score, and the area under the ROC curve (AUC). The numerical values of these four metrics were found to be 90.00%, 89.58%, 90.13%, 81.13%, and 0.9435, respectively. For a fully-supervised learning scheme (with 100% of the labeled training samples used for model training), the corresponding values were 90.50%, 91.67%, 90.13%, 82.25%, and 0.9457, respectively. When three professional anesthesiologists conducted comprehensive evaluation, the corresponding results were 91.00%, 91.67%, 90.79%, 83.26%, and 0.9497, respectively. It can be seen that the semi-supervised deep learning model trained by us with only 30% labeled samples can achieve a comparable effect with the fully supervised learning model, but the sample labeling cost is smaller. Our method can achieve a good balance between performance and cost. At the same time, the results of the semi-supervised model trained with only 30% labeled samples were very close to the performance of human experts.

Conclusions: To the best of our knowledge, our study is the first one to apply a semi-supervised deep-learning method in order to identify the difficulties of both mask ventilation and intubation. Our AI-based image analysis system can be used as an effective tool to identify patients with difficult airway conditions.

Clinical trial registration: ChiCTR2100049879 (URL: http://www.chictr.org.cn).

[*] Corresponding author.
E-mail address: zhangjiq@zzu.edu.cn (J. Zhang).

[1] Guangzhi Wang and Chenxi Li contributed equally to this work.

https://doi.org/10.1016/j.heliyon.2023.e15629
Received 13 October 2022; Received in revised form 15 April 2023; Accepted 17 April 2023
Available online 22 April 2023

1. Introduction

A difficult airway is defined by a spectrum of clinical situations in which a conventionally-trained anesthesiologist faces difficulties (or fails) in providing mask ventilation, performing trachea intubation, or both [1]. The incidence rate of mask ventilation difficulties ranges from 1.4% to 5.0%, while that of intubation difficulties ranges from 1.9% to 10% [2–7]. Unanticipated difficult airway conditions increase the risks of brain injury and mortality [8,9]. Furthermore, the treatment of such conditions requires specialized expertise and complicated procedures [10]. So, patients at risk of airway difficulties should be preoperatively identified in order to reduce the detrimental complications of these difficulties, and decrease the risk of anesthesia-induced mortality [9,11,12]. Nevertheless, predicting airway difficulties currently remains a challenging task, and the diagnostic accuracies of the existing prediction methods are still quite low [2,13]. Indeed, no satisfactory prediction performance for difficult airways can be achieved through common bedside screening tests [5]. In addition, comprehensive clinical evaluation methods for difficult airways (such as the modified LEMON criteria and the Simplified Airway Risk Index (SARI) model) cannot achieve the desired prediction performance, either [14,15].

The applications of artificial intelligence (AI) are expanding worldwide and its supporting role in medicine is increasingly being investigated [16]. At present, one of the key AI paradigms is deep learning. One significant advantage of deep learning is its ability to readily identify data patterns not typically recognized by human experts [16,17]. In recent years, deep learning has achieved remarkable outcomes in the detection and diagnosis of different clinical conditions, such as dermatopathy, diabetic retinopathy, oral squamous-cell carcinoma and even cancer types of unknown primary origins [17–20]. Patients with difficult intubation risks mostly manifest anatomical airway abnormalities, which anesthesiologists usually detect through naked-eye observations. Compared to human experts, the AI-based approaches appear to be better in identifying visual cues, and handling subtleties more objectively, accurately, and reliably [16,17].

To date, some research attempts have been made towards the development of image-based AI systems for difficult airway identification and management [21–24]. However, the existing AI-based methods for difficult airway management are still generally unsatisfactory due to several limitations such as the outdated AI algorithms, inadequate image acquisition standards, and low predictive accuracies. Moreover, the existing studies usually focused on intubation difficulties, but ignored mask ventilation difficulties.

In this paper, we introduce for the first time a new perioperative airway evaluation database. This database provides different vital signs and mobility metrics including the patient's natural appearance, mandibular mobility, maximum mouth opening, thyromental distance, Mallampati classification, neck movements, and chin-to-chest distance. Through this database, we develop a novel semi-supervised deep learning model which can be adequately trained for both types of difficult airway prediction. We believe the proposed SSL model could provide a new, reliable and practical method for perioperative prediction and evaluation of difficult airways.

2. Methods

2.1. Study design and participants

We conducted an observational study of 1000 patients who were scheduled for elective surgical procedures under general anesthesia at Henan Provincial People's Hospital, Henan, China. The study design was reviewed and approved on May 24, 2021 by the Institutional Ethics Committee (chaired by Prof. Feng-Min Shao) at that hospital (Protocol number: LS(2021–57)). A written informed consent was obtained for each participant. All patients participating in this study had consented for their images, clinical data and other relevant information to be published. Our study was registered in the Chinese Clinical Trial Registry, a primary registry of the International Clinical Trial Registry Platform, the World Health Organization (clinical trial registration number: ChiCTR2100049879; URL: http://www.chictr.org.cn). Our study complies with all pertinent regulations and the patients were enrolled in this study based on four inclusion criteria: (1) The patient was older than 18; (2) the patient had an elective surgery under general anesthesia; (3) the patient had a grade I–III according to the physical status classification of the American Society of Anesthesiologists (ASA); and (4) the patient voluntarily signed an informed consent form. Other patients were excluded based on three exclusion criteria: (1) patients with head or neck deformities, oral or maxillofacial deformities, absence of incisors, subglottic stenosis, severe maxillofacial trauma or cervical spondylosis; (2) patients with central nervous system diseases or psychiatric disorders; and (3) patients with communication disorders (language barriers, or severe visual or hearing impairment). Generally, according to the sampling distribution theory, the sample statistics asymptotically follow a normal distribution. Here, a 90% confidence interval was used with a sampling error of 3%. A sample size of 800 is generally needed for sound statistical inference [25]. As mentioned above, we collected data for 1,000 patients from August to October of 2021. Of the 1000 patients, 455 were male, and 545 were female, and their mean age was 62.8 years. The average BMI of all patients was 23.5. Patients had the American Society of Anesthesiologists Physical status (ASA PS) 1–3, with 14.3% having ASA PS 1, 69.3% having ASA PS 2, and 16.4% having ASA PS 3.

2.2. Study procedure

According to the study design, a written informed-consent form was obtained for each patient. Then, images were collected for each patient from 9 different viewpoints during a routine preoperative anesthesia procedure. The collected image set was divided into subsets in the ratio of 8:2 for training and testing the proposed model, respectively. For each test patient, the automated airway assessment results were concealed from the handling anesthesiologist. However, the anesthesiologist could freely evaluate the patient airways. As each patient entered the surgery room, routine monitoring was initiated for key vital signs including the electrocardiogram (ECG), non-invasive blood pressure (NBP), oxygen saturation (SpO_2), and end-tidal carbon dioxide concentration ($EtCO_2$). Anesthesia was induced by intravenously injecting midazolam (0.03 mg/kg), sufentanil (0.5–1.0 μg/kg), etomidate (0.3–0.6 mg/kg) and rocuronium (0.6–0.9 mg/kg). After administering these anesthetics, pure oxygen was supplied to each patient through a mask ventilation procedure, while the difficulty of this procedure was simultaneously evaluated. Moreover, for each patient, an anesthesiologist (with more than 5 years of clinical experience) performed a tracheal intubation procedure using a visual laryngoscope (after complete relaxation of the patient's muscles), observed the patient's laryngeal state, and then evaluated the intubation difficulty based on the Cormack–Lehane (CL) classification system [26]. In the CL system, Grade I indicates that the entire glottic aperture can be fully viewed; Grade II indicates that the vocal cords can be only partially viewed; Grade III reflects that only the epiglottis is viewable; and Grade IV represents the inability to view even the epiglottis. Airways with Grades I or II were regarded as normal airways, while those with Grades III or IV were labeled as difficult airways for tracheal intubation. Once a difficult intubation case was encountered, treatment guidelines were immediately followed with the help of a senior physician if necessary. In addition, patient safety may be further enhanced through using fiber bronchoscopes, employing light rods (to guide tracheal intubation), or constructing emergency surgical airways. If a patient has been confidently diagnosed with a difficult airway condition preoperatively, tracheal intubation should be performed without anesthesia under the guidance of a fiber bronchoscope or a light rod. If no definite CL classification could be determined for a certain case, that case would be directly marked as a difficult airway case in our database.

2.3. Image acquisition standard

In order to establish a perioperative airway assessment database, we collected nine face images of each patient as well as six difficult airway predictors: mandibular mobility, maximum mouth opening, the thyromental distance (TMD), Mallampati classification (MPC), neck movements, and the chin-to-chest distance [11]. These predictors account for classical and modern clinical evaluation methods of difficult airways. The setup and significance for each of the captured nine images of a patient are as follows:

- **Image 1:** A frontal image of the patient with the mouth open and the teeth closed at the same time to observe the dentition;
- **Image 2:** A frontal image of the patient protruding and closing the lower incisors to check whether the mandibular protrusion is limited;
- **Image 3:** A frontal image of the patient with the face at a neutral expression and the neck completely exposed to check for obesity and anatomical deformity signs (in combination with Image 6);

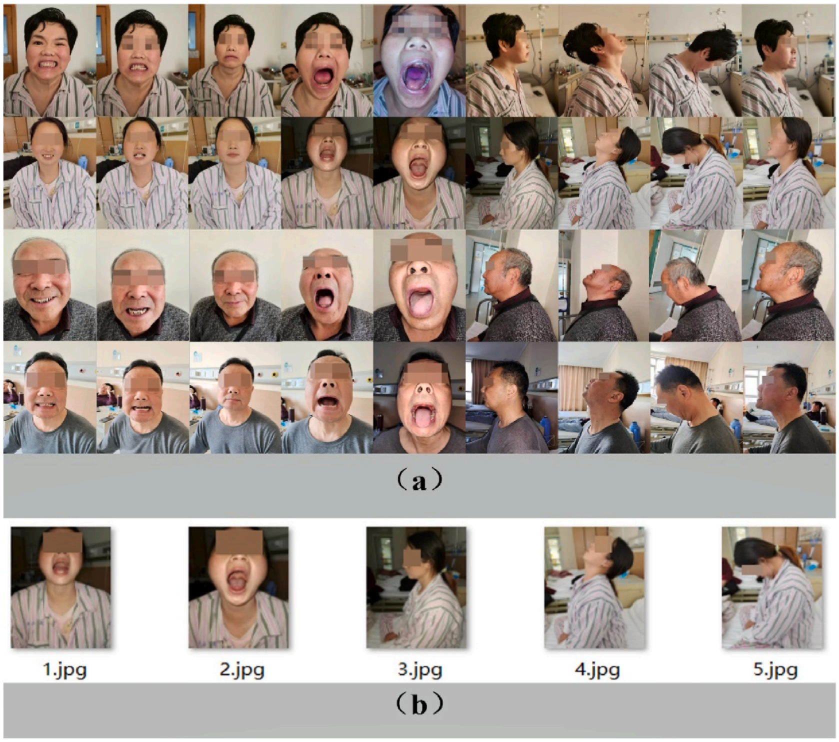


Fig. 1. (a) Sample images for some patients in the collected dataset. Each row corresponds to one patient, and 9 images are collected for each patient according to the image acquisition standard described in the text. **(b)** Five selected images with different angles. These images account for the patient's maximum mouth opening, Mallampati classification, neck length, neck circumference, neck movements, and thyromental distance.

- **Image 4:** A frontal image of the patient with the mouth maximally open to evaluate the degree of mouth opening;
- **Image 5:** A frontal image of the patient in an upright seated position, the head in a neutral position, the mouth open, and the tongue stuck out in order to observe the palatine arch, the uvula and the soft palate for Mallampati classification;
- **Image 6:** A side image of the patient in an upright seated position with the head in a neutral position and the neck completely exposed to check for obesity and anatomical deformity signs (in combination with Image 3);
- **Image 7:** A side image of the patient leaning back the neck as much as possible to observe the head and neck mobility;
- **Image 8:** A side image of the patient with maximum neck flexion to observe the head and neck mobility;
- **Image 9:** A side image of the patient in an upright seated position with the head in a neutral position, the lower incisors protruding, and the upper lip bitten for the occlusion test.

Fig. 1(a) shows some samples of the captured nine images from each patient. Meanwhile, parts around the mouth and the lips were blurred in these images in order to hide other facial features, improve the prediction accuracy, and protect their personal privacy. The collected images constitute the airway image dataset used in this work. Fig. 1 shows some samples of the collected facial images of the patients.

In our work, a semi-supervised deep learning (SSL) method is used for the first time to predict difficult airway conditions. The SSL framework reduces the manpower and resources needed for manual data labeling through effective exploitation of domain knowledge. Also, such a framework can lead to relatively high accuracies, and is thus quite applicable in medical image processing. However, for performance comparison of our semi-supervised learning method against baseline supervised learning methods, labels for all samples in our training dataset are needed. Unlabeled data samples can be obtained from our dataset by simply discarding the corresponding labels.

For manual data labeling, three anesthesiologists with more than five years of individual clinical experience examined and graded each of the captured patient images in terms of airway difficulty. Based on this expert evaluation, all patients were categorized into difficult-airway patients or normal-airway patients. Three anesthesiologists were recruited to assess the airway of each patient. Each expert conducted the assessment independently from other experts, and the assessment results were not revealed to the patients. Finally, after excluding unusable samples according to the aforementioned inclusion/exclusion criteria, our dataset had a total of 1,000 patients (including 758 non-difficult-airway patients and 242 difficult-airway patients).

2.4. *Proposed approach*

In this paper, we present the proposed semi-supervised deep learning method for predicting difficult airway conditions from multi-

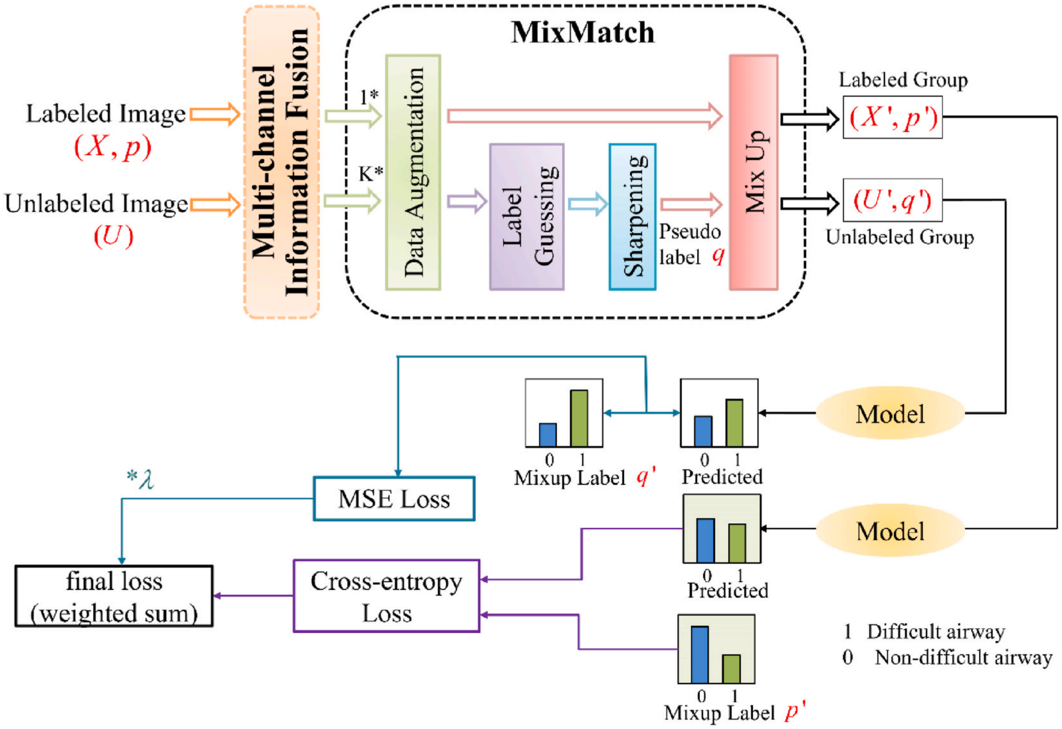


Fig. 2. The overall framework of the semi-supervised deep learning method for difficult airway prediction. This framework has three main modules: multi-channel information fusion, the MixMatch semi-supervised learning algorithm, and the ResNet neural network model. The MixMatch module consists of four parts: data augmentation, label guessing, sharpening, and MixUp (See the text for details).

view images. Specifically, we use a multi-channel information fusion approach to fuse multi-view patient images, and then use the MixMatch [27] semi-supervised deep learning method along with the ResNet18 [28] architecture as the backbone network for airway classification. Fig. 2 shows the workflow of the proposed approach.

2.4.1. *Data preprocessing*

In order to simplify the difficult airway prediction model, we used only five out of nine patient images (namely, the fourth, fifth, sixth, seventh and eighth images) for model construction (See Fig. 1(b)). The selected images highlight six key pieces of patient information: the maximum mouth opening, the Mallampati classification, the neck length, the neck circumference, neck motion, and the parathyroid distance. Moreover, the number of the non-difficult-airway patients in the collected dataset was approximately three times that of the difficult-airway patients. To alleviate this significant class imbalance, we used data augmentation techniques in order to increase the numbers of the training samples of the non-difficult-airway cases and the difficult-airway cases by factors of 2 and 6, respectively. The employed data augmentation techniques were specifically the following: horizontal flipping, random rotations within 10°, scaling, as well as adjustment of brightness, contrast, saturation, and hue. After data augmentation, there was no significant difference in the numbers of difficult and non-difficult airway patients.

2.4.2. *Multi-channel information fusion*

The existing deep learning methods for difficult airway prediction are typically based on single-view image data (with no multi-channel data), and thus these methods aren't highly reliable. In our work, we adopt a multi-channel information fusion method to fuse the aforementioned five image views of a patient's airway. This fusion process produces a higher-dimensional image that captures the patient's maximum mouth opening, the Mallampati classification, the neck length, the neck circumference, neck movements, the thyromental distance and other characteristic information. The fused information is quite comprehensive and the potential for better model reliability is higher. Since a color image has three channels (R, G, and B), the fused image of a single patient has a size of 224 × 224 × 15.

2.4.3. *MixMatch*

MixMatch is a "holistic" learning approach which incorporates ideas and components from the dominant paradigms of semi-supervised learning. Table 1 shows the full MixMatch algorithm. Fig. 2 shows the flow of the semi-supervised deep learning algorithm MixMatch for difficult airway prediction. The total loss is calculated using the following set of equations:

$$X', U' = MixMatch(X, U, T, K, \alpha) \tag{1}$$

$$L_X = \frac{1}{|X'|}\sum_{x,p\in X'} H\left(p', P_{\mathrm{mod}\ el}(y|x;\theta)\right) \tag{2}$$

$$L_U = \frac{1}{l|U'|}\sum_{u,q\in U'} \left\| q' - P_{\mathrm{mod}\ el}(y|u;\theta) \right\|_2^2 \tag{3}$$

Table 1
Algorithm 1.

Algorithm 1 MixMatch takes a batch of labeled data X and a batch of unlabeled data U and produces a collection X' (resp. U') of processed labeled examples (resp. unlabeled with guessed labels).

```
1:Input: Batch of labeled examples and their one-hot labels $X = ((x_b, p_b); b \in (1,...,B))$, batch of unlabeled examples $U = (u_b; b \in (1,...,B))$, sharpening temperature
    $T$, number of augmentations $K$, Beta distribution parameter $\alpha$ for MixUp.
    2: for $b = 1$ to $B$ do
    3: $\hat{x}_b = Augment(x_b)$//Apply data augmentation to $x_b$
    4: for $k = 1$ to $K$ do
    5: $\hat{u}_{b,k} = Augment(u_b)$//Apply $k^{th}$ round of data augmentation to $u_b$
    6: end for 7: $\overline{q_b} = \frac{1}{K}\sum_k P_{\mathrm{mod}\ el}(y|\hat{u}_{b,k};\theta)$//Label Guessing
    8: $q_b = Sharpen(\overline{q_b}, T)$//Temperature sharpening
    9: end for 10: $\hat{X} = ((\hat{x}_b, p_b); b \in (1,...,B))$//Augmented labeled examples and their labels
    11: $\hat{U} = ((\hat{u}_{b,k}, q_b); b \in (1,...,B), k \in (1,...,K))$
    //Augmented unlabeled examples, guessed labels
    12: $W = Shuffle(Concat(\hat{X}, \hat{U}))$//Combine and shuffle labeled and unlabeled data
    13: $X' = (MixUp(\hat{X}_i, W_i); i \in (1,...,|\hat{X}|))$
    //Apply MixUp to labeled data and entries from W
    14: $U' = (MixUp(\hat{U}_i, W_{i+|\hat{X}|}); i \in (1,...,|\hat{U}|))$
    //Apply MixUp to unlabeled data and the rest of W
    15: return $X', U'$
```

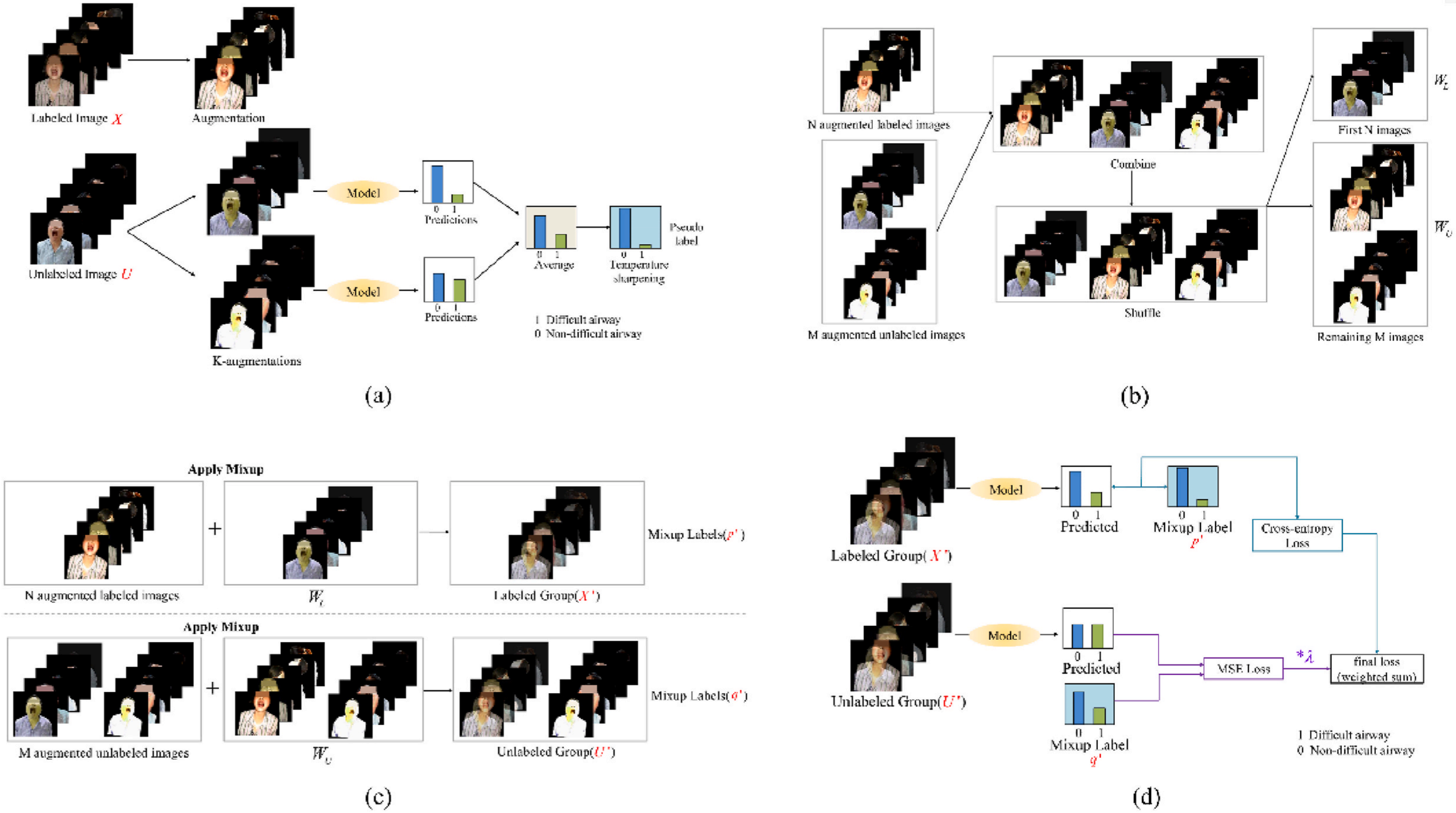


Fig. 3. Steps of the MixMatch module. **(a) Data augmentation, label guessing, and sharpening.** Each of the labeled samples *X* and the unlabeled samples *U* were treated with one and *K* data augmentation rounds, respectively. The results of the *K* data augmentation rounds for each unlabeled sample were predicted by the model, and then averaged. Finally, the classification entropy was modulated by temperature sharpening to generate a “guess” pseudo-label for the unlabeled sample. **(b) Step 1 of MixUp:** Combine the enhanced labeled and unlabeled samples and rearrange them after shuffling. **(c) Step 2 of MixUp:** Match the samples of the original order with the samples of the shuffled order, and then apply MixUp separately. **(d) Loss calculation.** For the labeled samples, calculate the cross-entropy loss between the model-predicted results and the true labels P' after MixUp. For the unlabeled samples, calculate the mean-square error (MSE) loss between the model-predicted results and the pseudo-labels q' after MixUp. Finally, combining these two losses with different weights constitutes the total loss function.

In Eq. (1), X denotes a batch of labeled data with labels p, while U denotes a batch of unlabeled data. The labeled data (X', p') and the unlabeled data (U', q') are generated after MixMatch, where $q^{'}$ denotes the "guess" labels of the unlabeled data, while T, K, α are hyperparameters. In Eq. (2), L_X represents the loss of the labeled data, $|X^{'}|$ represents the number of the labeled examples, and $H(p^{'}, P_{\mathrm{mod}\ el})$ is the cross-entropy between the ground-truth sample label $p^{'}$ and the model prediction $P_{\mathrm{mod}\ el}$. In Eq. (3), L_U represents the loss of the unlabeled data, l represents the number of categories, $|U^{'}|$ represents the number of the unlabeled examples, and $\|.\|_2^2$ denotes the L_2 norm, where L_2 regularization [29,30] is used. In Eq. (4), λ_U is a weight coefficient, which represents the proportion of the unlabeled data loss to the total loss. The labeled data loss L_X and the unlabeled data loss L_U are combined with different weights to form the total loss function L.

As shown in Fig. 2, the MixMatch module consists of four parts: data augmentation, label guessing, sharpening, and mixup. The complete MixMatch algorithm is shown in Algorithm 1. The MixMatch workflow is graphically shown in Fig. 3(a–c), while Fig. 3(d) gives a schematic diagram for the loss function calculations.

① Data augmentation

We perform data augmentation on both labeled and unlabeled data samples. Data augmentation applies input transformations assumed to leave class semantics unaffected. For example, in image classification, it is common to elastically deform or add noise to an input image, which can dramatically change the pixel content of an image without altering its label [31–33]. Roughly speaking, this can artificially expand the size of a training set by generating a near-infinite stream of new, modified data. Also increasing the type of samples and enhancing the generalisation performance of the model. For each x_b in the batch of the labeled data X, we perform data augmentation once to generate $\overset{\wedge}{x_b} = Augment(x_b)$ (Algorithm 1, line 3). For each u_b in the batch of the unlabeled data U, we perform data augmentation K times to generate $\overset{\wedge}{u_{b,k}} = Augment(u_b), k \in (1, ..., K)$ (Algorithm 1, line 5).

② Label guessing & sharpening

For each unlabeled example in U, we use the model prediction result to generate a "guessed" label, which is subsequently used to calculate the loss of the unlabeled data [34–36]. The model is used to predict the class distribution of u_b's K data augmentation results $\overset{\wedge}{u_b}$ and hence obtain the average value (Algorithm 1, line 7):

$$\overline{q_b} = \frac{1}{K}\sum_k P_{\mathrm{mod}\ el}\left(y \middle| \overset{\wedge}{u_{b,k}}; \theta\right) \tag{5}$$

In Eq. (5), $\overline{q_b}$ is the average of the model's predicted class distributions across all the K augmentations of u_b. Based on entropy minimization [37] in semi-supervised learning, we added a sharpening step in the process of generating "guess" labels. Specifically, after obtaining the average $\overline{q_b}$, we use the *Sharpen* function to decrease the entropy loss, and we employ a parameter T to control this loss. The *Sharpen* function is mathematically expressed as follows

$$Sharpen(p, T)_i : = {p_i}^{\frac{1}{T}} \Bigg/ \sum_{j=1}^{L} {p_j}^{\frac{1}{T}} \tag{6}$$

In Eq. (6), p denotes the class distribution predicted by the model (in the MixMatch module, p indicates the average class prediction $\overline{q_b}$ of the K augmentation results, as shown in Algorithm 1, line 8), and T is a hyperparameter that can be used to adjust the entropy loss. In the MixMatch module, reducing the parameter T can encourage the model to make low-entropy predictions.

③ MixUp

Actually, MixUp [38] is an unconventional data augmentation method. In the aforementioned semi-supervised learning algorithm, the labeled and unlabeled data samples are completely separate. Here, MixUp is used to fuse such labeled and unlabeled samples, where the data generated by MixUp carries information about these two types of samples. We use here a fine-tuned variant of the original MixUp method. For two labeled examples (x_1, p_1) and (x_2, p_2), we use the following formulas to calculate the MixUp-generated data $(x^{'}, p^{'})$.

$$\lambda \sim Beta(\alpha, \alpha), \tag{7}$$

$$\lambda^{'} = \max(\lambda, 1 - \lambda), \tag{8}$$

$$x' = \lambda' x_1 + (1 - \lambda')x_2, \tag{9}$$

$$p' = \lambda' p_1 + (1 - \lambda')p_2, \tag{10}$$

In Eq. (7), α is a hyperparameter. While $\lambda^{'} = \lambda$ in the original MixUp method, our fine-tuned MixUp method uses Eq. (8) to calculate $\lambda^{'}$, such that $x^{'}$ is closer to x_1 instead of x_2. In Eq. (9) and Eq. (10), $x^{'}$ is the new sample after MixUp, and $p^{'}$ is the new label after MixUp.

In order to use MixUp to fuse labeled and unlabeled examples, we collect all the augmented labeled examples into the collection $\hat{X}$ (Algorithm 1, line 10), and the unlabeled examples with "guess" labels into the collection $\hat{U}$ (Algorithm 1, line 11), as shown in Eq. (11) and Eq. (12).

$$\hat{X} = ((\hat{x}_b, p_b); b \in (1, ..., B)), \tag{11}$$

$$\hat{U} = \left((\hat{u}_{b,k}, q_b); b \in (1, ..., B), k \in (1, ..., K)\right) \tag{12}$$

Then, we combine and shuffle the collections $\hat{X}$ and $\hat{U}$ to form a collection W, which will be used as the data source for MixUp (Algorithm 1, line 12). For each example-label pair in $\hat{X}$, we calculate $MixUp(\widehat{X}_i, W_i); i \in (1, ..., |\widehat{X}|)$ and add the result to the collection X' (Algorithm 1, line 13). From Eqs. (8) and (9), we know that the examples in and $\hat{X}$ are highly similar. For each example-label pair in $\hat{U}$, we apply MixUp with the remaining example-label pairs in W and add the result to the collection U' (Algorithm 1, line 14). The MixMatch algorithm converts X to X', where X' is a collection of labeled examples obtained with one application of data augmentation and MixUp. Similarly, U is converted to U', where each unlabeled example in U' has undergone K rounds of data augmentation and one MixUp round, and has a corresponding "guess" label.

3. Experimental setup, results and analysis

3.1. Experimental setup

3.1.1. Settings

As mentioned before, our dataset contains samples for a total of 1000 patients, including 758 non-difficult-airway patients and 242 difficult-airway patients. For model training, we used 80% of the data (800 patients including 606 non-difficult-airway patients and 194 difficult-airway patients). For model testing, we used the remaining 20% of the data (200 patients with 152 non-difficult-airway patients and 48 difficult-airway patients). In order to avoid the dependence of the results on a particular division of the data into training and testing subsets, we followed a five-fold cross-validation scheme as shown in Fig. 4. In this validation scheme, the dataset is divided evenly into five folds, where each fold (20% of the data) is used in turn for testing while the remaining four folds (80% of the data) are used for training. According to this scheme, five training-testing data partitions (Dataset 1–Dataset 5) are produced with the same training-testing ratios but with completely different testing samples and partially different training samples. The five partitions are independently used to produce five different classification models. The test results for these five models are obtained and averaged to get the overall test results.

To alleviate the class imbalance in the training data, we employed data augmentation in the preprocessing stage. The details of the data division among the two classes are given in Table 2.

In order to ensure better model convergence, improve experimental efficiency, and reduce over-fitting, we used a transfer learning approach to retrain a ResNet18 architecture (pre-trained on ImageNet [28]) as our baseline model. For the semi-supervised learning method, we used a mixed loss function composed of a labeled data loss and an unlabeled data loss. For the supervised learning method, a cross-entropy loss was used. For all experiments, we used 90 epochs of training, an Adam optimizer, a batch size of 32, an initial learning rate of 0.001, and a learning rate adjustment to 0.0001 after two-thirds of the epochs.

All experiments were run within Anaconda 4.10.1 on a Windows 10 operating system, with Python 3.7.10, PyTorch 1.8.1 and other packages. We used a graphics card of NVIDIA GeForce RTX 3070 with an 8-GB memory capacity, a GDDR8 memory type, and a 14000-MHz frequency.

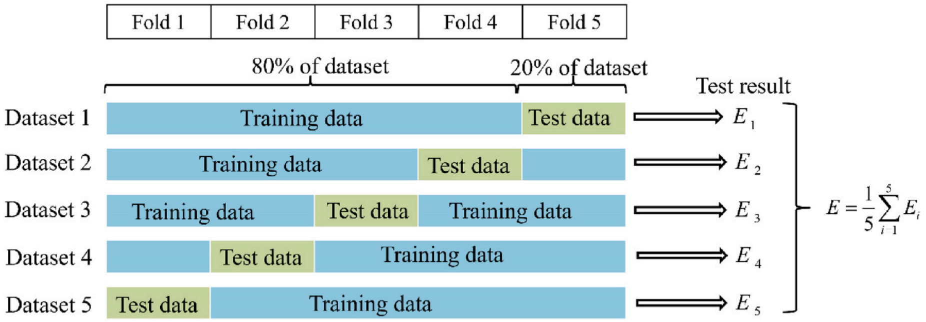


Fig. 4. Five-fold cross-validation. The dataset is divided into five folds, where each fold is used in turn as the test set while the other four are used as the training set. Thus, five different data combinations (Dataset 1–Dataset 5) are generated, and five corresponding classification models are trained and tested to get the results E1–E5. The average of these five test results is the final test result E.

Table 2
Dataset division.

	Training data		Training data after data expansion		Test data	
	Non-difficult	Difficult	Non-difficult	Difficult	Non-difficult	Difficult
Dataset 1	606	194	1212	1164	152	48
Dataset 2	606	194	1212	1164	152	48
Dataset 3	606	194	1212	1164	152	48
Dataset 4	606	194	1212	1164	152	48
Dataset 5	606	194	1212	1164	152	48

3.1.2. Performance indicators

In this study, we used the accuracy, sensitivity, specificity, precision, and F1-score as overall performance indicators. These indicators are respectively defined as follows:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{13}$$

$$Sen = \frac{TP}{TP + FN} \tag{14}$$

$$Spe = \frac{TN}{FP + TN} \tag{15}$$

$$Pre = \frac{TP}{FP + TP} \tag{16}$$

$$F_1 = \frac{2 \times Pre \times Sen}{Pre + Sen} \tag{17}$$

In Eqs. (13) to (16), TP, TN, FP, FN denote the true-positive, true-negative, false-positive and false-negative, respectively. The result of Eq. (17) is calculated using the results of Eq. (14) and Eq. (16).

For medical image classification problems, another key performance evaluation method is the receiver operating characteristic (ROC) curve analysis and the associated area under the ROC curve (AUC). Indeed, the ROC curve represents a comprehensive indicator of binary classification performance, where the curve continuously traces all sensitivity-specificity pairs, and each ROC point reflects the susceptibility to the same signal stimulus. The AUC value is generally between 0.5 and 1, and is used to judge the overall model performance. The closer the AUC value to 1, the better the model performance.

3.2. Experimental results

First of all, we investigated the classification performance with different types of backbone networks, namely, DenseNet121 [39], GoogLeNet [40], InceptionV3 [41], MobileNetV2 [26], ShuffleNetV2 [42], VGG16 [43], and ResNet18 [44]. For the semi-supervised deep-learning algorithms, we used each backbone network with the MixMatch method. The parameter settings of the different variants were kept consistent. From Table 3 and Fig. 5, we can observe that the best experimental results are obtained with a ResNet18 backbone network, and 30% of the training samples being labeled (and 70% of the training samples being unlabeled).

For our dataset, we evaluated the effectiveness of the MixMatch method through comparison with several semi-supervised learning methods: Π-Model [34], Mean Teacher [36], virtual adversarial training (VAT) [45], and Pseudo-Label [46]. For a fair comparison, all methods were implemented and compared with the same backbone network, ResNet18. Table 4 shows the performance outcomes of these methods with 30% (410) labeled training samples. The fully-supervised model trained with 100% (1368) labeled samples is used to set an upper bound on performance, while the fully-supervised model trained with 30% (410) labeled samples is used to set the baseline performance. The assessments of the three experienced anesthesiologists were combined to get the overall ground-truth

Table 3
Comparison of the metrics of our approach when using different backbone networks.

Method	Percentage		Metrics				
	Labeled	Unlabeled	Acc	Sen	Spe	F1	AUC
DenseNet121	30%	70%	86.50%	79.17%	88.82%	73.79%	0.8936
GoogLeNet	30%	70%	79.50%	72.92%	81.58%	63.06%	0.8136
InceptionV3	30%	70%	84.00%	81.25%	84.87%	70.91%	0.8588
MobileNetV2	30%	70%	85.00%	83.33%	85.53%	72.73%	0.8782
ShuffleNetV2	30%	70%	83.00%	83.33%	82.89%	70.18%	0.8495
VGG16	30%	70%	87.50%	89.58%	86.84%	77.47%	0.9035
ResNet18	30%	70%	**90.00%**	**89.58%**	**90.13%**	**81.13%**	**0.9435**

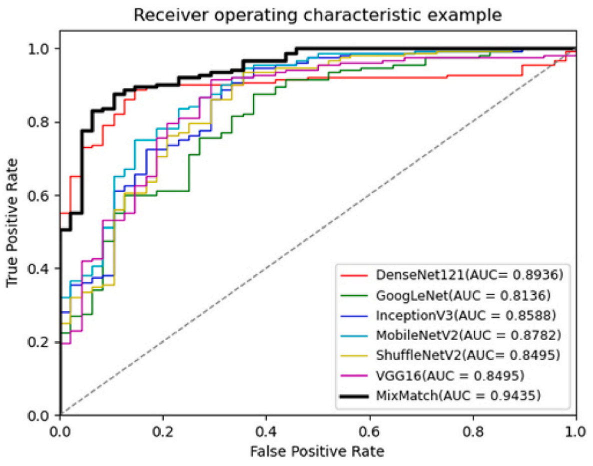


Fig. 5. Comparison of the ROC curves for our approach in combination with different backbone networks.

Table 4
Comparison of experimental results between different semi-supervised deep learning methods.

Method	Percentage		Metrics				
	Labeled	Unlabeled	Acc	Sen	Spe	F1	AUC
doctor	-	-	**91.00%**	**91.67%**	**90.79%**	**83.26%**	**0.9497**
Upper Bound	100%	0	**90.50%**	**91.67%**	**90.13%**	**82.25%**	**0.9457**
Baseline	30%	0	85.50%	81.25%	86.84%	72.90%	0.8850
Π-Model [34]	30%	70%	87.00%	83.33%	88.16%	75.47%	0.9002
Mean Teacher [36]	30%	70%	88.50%	85.42%	89.47%	78.10%	0.9110
VAT [45]	30%	70%	89.00%	87.50%	89.47%	79.24%	0.9127
Pseudo-Labe [46]	30%	70%	86.00%	81.25%	87.50%	73.58%	0.8936
MixMatch	30%	70%	**90.00%**	**89.58%**	**90.13%**	**81.13%**	**0.9435**

expert ratings. As we can see, all semi-supervised learning models gained some improvement in terms of all performance metrics compared to the baseline model. In particular, the MixMatch-based method gained better performance than the other semi-supervised learning methods. The upper-bound accuracy (with all training samples being labeled) was only 0.5% lower and the AUC was only 0.004 lower compared to the performance indicators of the human experts. However, the fully-supervised method (with the upper-bound performance) requires labeling all of the training samples, and is thus quite time-consuming and laborious. In comparison, the MixMatch-based semi-supervised learning method demonstrated a good balance between performance and cost, where the obtained accuracy and AUC were respectively just 1% and 0.0062 lower than the indicators of the human experts, with only 30% of the training samples being labeled. Fig. 6 shows the ROC curves obtained for various methods. Clearly, the performance of our MixMatch-based method is better compared to other semi-supervised learning methods, and is very close to the upper-bound

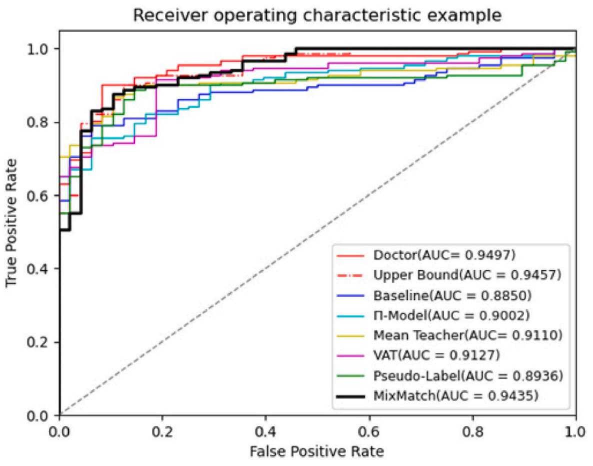


Fig. 6. Comparison of the ROC curves associated with different semi-supervised deep learning methods.

Table 5
Comparison of metrics of MixMatch-based semi-supervised learning methods for different input cases.

Input	Percentage		Metrics				
	Labeled	Unlabeled	Acc	Sen	Spe	F1	AUC
1.jpg	30%	70%	78.50%	77.08%	78.95%	63.24%	0.7841
2.jpg	30%	70%	83.50%	83.33%	83.55%	70.79%	0.8527
3.jpg	30%	70%	72.50%	72.92%	72.37%	56.00%	0.7401
4.jpg	30%	70%	76.00%	75.00%	76.32%	60.00%	0.7688
5.jpg	30%	70%	75.00%	72.92%	75.66%	58.33%	0.7538
1.jpg~5.jpg	30%	70%	**90.00%**	**89.58%**	**90.13%**	**81.13%**	**0.9435**

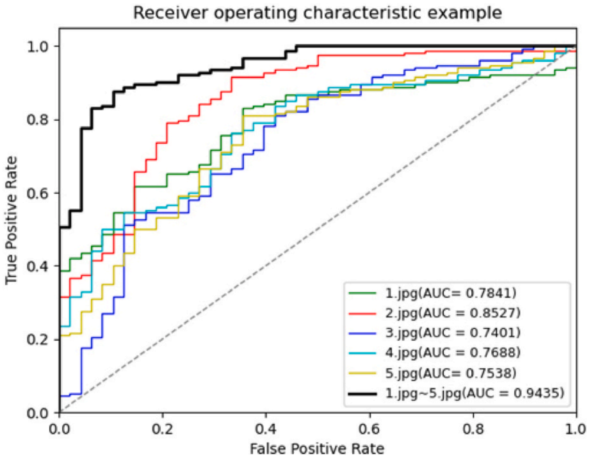


Fig. 7. Comparison of the ROC curves of the MixMatch-based semi-supervised learning methods for different input cases.

Table 6
Quantitative evaluation of our method was performed on the constructed dataset under different percentages of labeled data.

Method	Percentage		Metrics				
	Labeled	Unlabeled	Acc	Sen	Spe	F1	AUC
doctor	–	–	91.00%	91.67%	90.79%	83.26%	0.9497
Upper Bound	100%	0	90.50%	91.67%	90.13%	82.25%	0.9457
Baseline	5%	0	81.50%	79.17%	82.24%	67.26%	0.8230
MixMatch	5%	95%	84.00%	83.33%	84.21%	71.43%	0.8699
Baseline	10%	0	82.50%	81.25%	82.89%	69.03%	0.8441
MixMatch	10%	90%	86.00%	83.33%	86.84%	74.07%	0.8845
Baseline	20%	0	84.00%	83.33%	84.21%	71.43%	0.8669
MixMatch	20%	80%	88.50%	87.50%	88.82%	78.51%	0.9032
Baseline	30%	0	85.50%	81.25%	86.84%	72.90%	0.8850
MixMatch	30%	70%	90.00%	89.58%	90.13%	81.13%	0.9435

performance and the human expert performance.

To verify the effectiveness of our multichannel information-fusion method, we compared the fused-input multi-image multi-angle model against the direct-input single-image single-angle model. As the results in Table 5 and Fig. 7 show, the fusion of the 5 images produced the best results, with all the performance metrics higher than the corresponding ones for the single-image model. This demonstrates the feasibility of the multichannel information-fusion approach, which accounts for the variability in patient characteristics, captures more comprehensive information, and exhibits higher reliability .

Furthermore, we investigated the impact of different percentages of training data labeling on the performance of our semi-supervised learning method. As Table 6 shows, our method achieves consistent improvements over the fully-supervised baseline model with labeling percentages of 5%, 10%, 20% and 30%. In addition, with a 30% labeling percentage, our method achieves an AUC of 0.9435, a sensitivity of 89.58%, and a specificity of 90.13%. In general, these results demonstrate the effectiveness of our approach in exploiting the unlabeled data for achieving performance gains.

4. Discussion

Our work introduces the first semi-supervised deep-learning method for identifying difficulties of both mask ventilation and intubation. In this work, we established a non-invasive, fast, and easy-to-use AI model for predicting difficult airways, Fig. 8 shows the application interface for difficult airway detection. Our model can be trained without a large number of labeled samples, and can greatly reduce the cost of medical expert labeling. Also, compared to existing bedside screening tests and systematic evaluation methods [11,24], our model demonstrated superior performance in terms of several performance indicators (such as the accuracy, AUC, sensitivity, and specificity). In addition, our AI model could more accurately differentiate between normal-airway patients and patients with risks of difficult intubation or difficult mask ventilation. Accordingly, the proposed method is clinically applicable with relatively low misdiagnosis rate, and relatively high reliability. Indeed, our model is accessible and useful for clinicians (especially young ones who lack clinical experience), and can actually help identify patients with risks of difficult intubation during preoperative anesthesia procedures.

With the emergence of modern anesthesiology, airway management of patients under anesthesia has been a key issue towards achieving optimal oxygenation and ventilation. Most anesthesiologists frequently experience difficulties with airway management, and this brings great distress to the patients and poses treatment challenges. After a general anesthesia procedure, the patient loses the ability to breathe spontaneously, and the anesthesiologist shall usually re-provide adequate oxygen supply to the patient in about 1 min through mechanical ventilation. This temporary asphyxia doesn't usually cause any harm to the patient. However, for patients with difficult airways, ventilation cannot be quickly reestablished, and repeated intubation (or other procedures) could potentially damage the teeth, oral mucosa, and vocal cords [47]. Additionally, prolonged hypoxia can lead to serious organ damage, such as anoxic brain injury, myocardial ischemia, and even death [8,9].

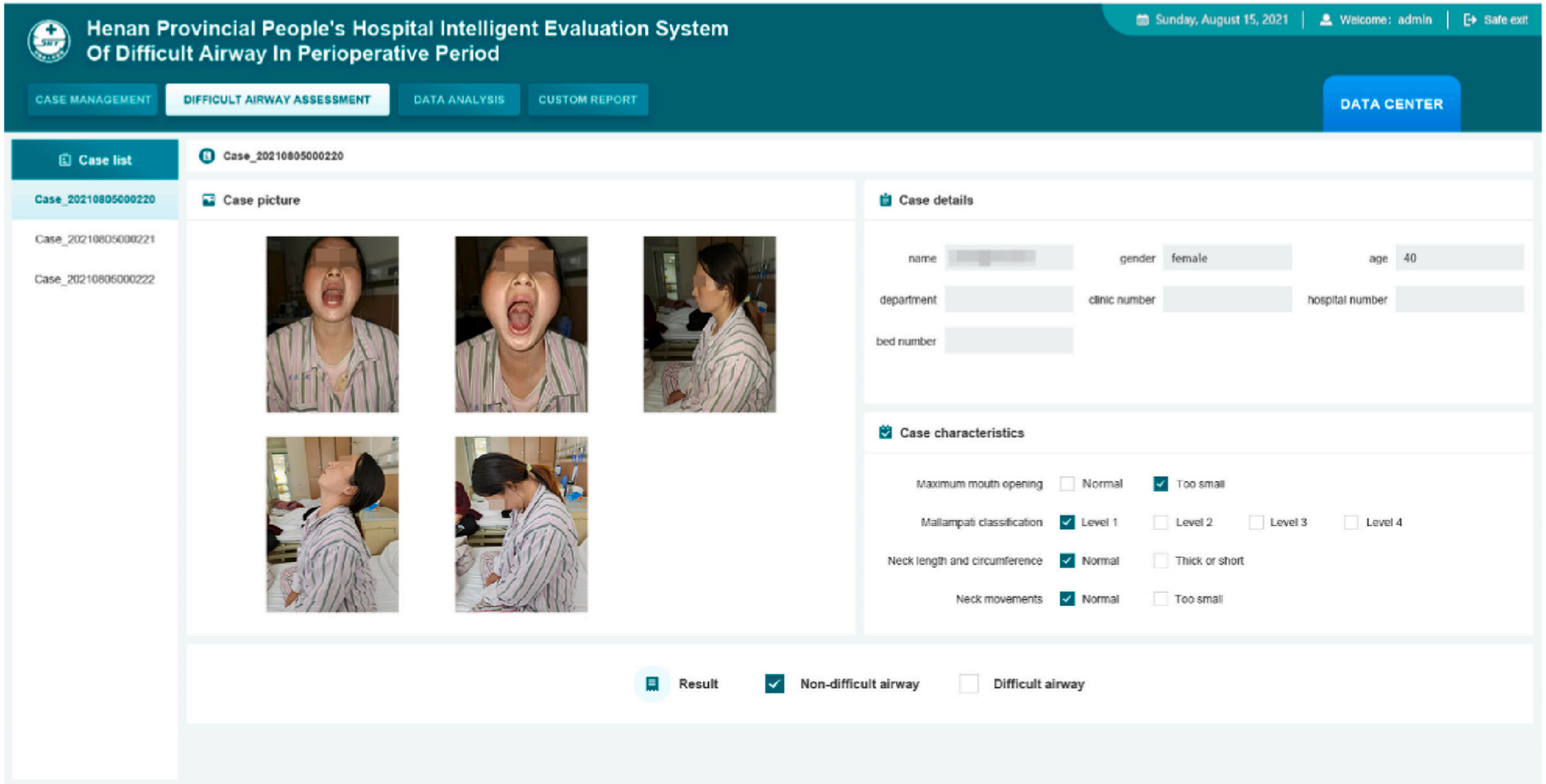
Therefore, the 4th National Audit Project (NAP4) of the Royal College of Anesthetists and the Difficult Airway Society as well as other major national anesthesia societies recommend preoperative airway assessment [1,8]. Not only can such an assessment inform patients about suspected airway difficulties and intubation-related risks, but also alert anesthesiologists to develop adequately-tailored preoperative anesthetic procedures (such as abandoning routine rapid continuous drug injection, utilizing slow-induction anesthesia with spontaneous respiration, or awake tracheal intubation recommended by most guidelines) [1,48]. Such procedures reduce the possibility of serious perioperative complications caused by inadequate preparation or anesthetic regimens. In addition, accurate preoperative prediction of difficult airway conditions could effectively help with avoiding difficult airway management failures caused by other human factors (such as inadequate teamwork, failure to timely seek help from superior medical experts, and improper preparation of the endotracheal tube and medications) [12,49,50]. However, no indicator or method can still accurately predict difficult airway conditions. Indeed, no satisfactory prediction performance could be achieved based on a patient's thyromental distance, chin-to-chest distance, maximum mouth opening and other indicators. Moreover, little improvements have been achieved through slightly complex clinical tests (such as the modified Mallampati classification and the upper lip bite test), or systematically-used multivariable models for airway assessment based on clinical experience (such as SARI, the Wilson risk score, and the LEMON method) [14,15,51].

Recently, AI techniques (especially deep learning ones) have been widely employed in multiple medical fields in order to help make better diagnosis, prognosis, and therapeutic decisions. Noticeably, the AI-based techniques for medical image analysis have been able to reach and even exceed the clinical performance of the human medical experts on certain clinical and diagnostic tasks (such as medical film reading) [52,53]. In fact, the human expert assessment of difficult airways is mostly based on visual information (with little attention to medical history). So, we believe that the AI-based predicative analytics can simplify the diagnosis of difficult airways. In fact, numerous AI-based methods have been proposed for predicting intubation difficulties. Hayasaka et al. established a deep learning model for difficult intubation classification using 16 patient head images [24]. However, just like other studies on difficult airway predication, the predication performance was limited by the small sample size (large datasets are critical for boosting the diagnosis performance) [20,54]. Therefore, considering the importance of the sample size on the predicative accuracy, we collected clinical information for 1000 patients in our study. In order to further increase the diversity of samples and improve the generalization performance of the model, we conducted data augmentation on the original data and used more diversified samples to train our AI model. In addition, to the best of our knowledge, we are the first to independently investigate the prediction of the mask ventilation difficulties. Despite the fact that very few cases of mask ventilation difficulties are available in our study, we believe that our study will be helpful to more comprehensively understand the risk factors of difficult airways.

In this study, we used a multi-channel information fusion method to fuse multiple images of different parts and different angles of each patient. Compared with the single-image-input neural networks in the literature [24], we feed our network with higher-dimensional information generated by multi-channel image fusion. The fused information takes into account the patient's maximum mouth opening, MPC, neck length, neck circumference, neck movements, TMD, and other key indicators. This information is more comprehensive and their reliability is relatively higher.

We also used a semi-supervised deep learning method to solve the difficult airway prediction problem for the first time. Compared with an existing supervised learning algorithm, we achieved comparable classification performance with only a small amount of labeled data. While reducing the demand for human and material resources, our method can also ensure good classification accuracy, and greatly improved efficiency.

We believe that the image acquisition location and angle are vital for the success of the AI-based predictions. Also, the success of our AI-based method does not mean that the clinical experience of the anesthesiologists is no longer important. No matter how science develops, the difficult airway predictors traditionally developed by anesthesiologists still have some clinical relevance and value. A wide range of bedside screening tests was utilized in our study, because just imaging a patient's visual appearance cannot well reflect the internal airway conditions, nor confer an advantage for AI-based image classification. This is also the part where our study stands out compared to other studies. During the pre-operative anesthesia visits, most patients sit or stand to communicate with their anesthesiologists, except for special physical reasons. Therefore, all patient images in our study were taken while the patients were



Henan Provincial People's Hospital Intelligent Evaluation System Of Difficult Airway In Perioperative Period

Case_20210805000220

name	[redacted]	gender	female	age	40
department		clinic number		hospital number	
bed number					

Maximum mouth opening	□ Normal	☑ Too small		
Mallampati classification	☑ Level 1	□ Level 2	□ Level 3	□ Level 4
Neck length and circumference	☑ Normal	□ Thick or short		
Neck movements	☑ Normal	□ Too small		

Result ☑ Non-difficult airway □ Difficult airway

Fig. 8. An application interface for difficult airway detection.

seated. Few previous studies claimed that the diagnostic performance with the supine position is better than that with the seated position [55]. However, the existing evidence is not sufficient enough to support this claim.

The direct laryngoscope is generally of low cost and thus has been more frequently used in previous studies than the visual laryngoscope. However, an increasing clinical evidence has demonstrated that the visual laryngoscope can reduce the intubation frequency, widen the visual field, and reduce the intubation-caused airway injuries [56]. Furthermore, the application of the visual laryngoscope can reduce the hospitalization cost, the length of stay, the risk of complications, and the risk of admission to an intensive care unit (ICU) [57]. We believe that the visual laryngoscope would be more widely used in the future as the main intubation tool in clinical anesthesia. Therefore, in this study, we chose the visual laryngoscope as our preferred intubation tool, and the exposure field of this device was used as the basis for difficult airway assessment.

5. Limitations of the study

However, our work has a few limitations. Firstly, this study is a single-center study with a potential sample bias. Also, in our study, the incidence of the difficult airway cases in the Henan province is not high. However, it is worth noticing that the number of patients with oral cancer is increasing in southern China due to betel nut chewing. Although there is no evidence of reaching an epidemic level, the spread of oral cancer may lead to higher prevalence of difficult airway cases. Secondly, compared to other AI methods with more than 10 thousand images of computed tomography (CT) or magnetic resonance imaging (MRI), the number of cases and images we studied is still relatively small. Finally, in clinical practice, in addition to visual indicators, the assessment of difficult airways by senior anesthesiologists will also be assisted by their subjective expertise and the patient's past medical history. Our AI-based method does not account for these pieces of information, and this represents a drawback of our work.

In view of the above shortcomings, we plan to carry out a multi-center study in collaboration with several hospitals of diverse geographical locations in China. This would enable the establishment of a larger database with more clinical images. At the same time, we have already begun the deployment of our AI model into a mobile phone application, in order to help clinical anesthesiologists use AI technologies to evaluate patient's airway conditions anytime and anywhere. Furthermore, in order to boost the diagnostic accuracy of our AI system, we are in the process of augmenting the system with manual input options for relevant factors (such as obesity, snoring and other medical history information).

Funding

Young and Middle-aged Leading Talents Fund in Health Science and Technology Innovation of Henan Province, China (YXKC2020001).

Provincial Department Co-Construction Project of Medical Science and Technology Project of Henan Province, China (SB201901090).

Collaborative Innovation Major Project of Zhengzhou (20XTZX05015).

Production notes

Author contribution statement

Guangzhi Wang: Conceived and designed the experiments; Performed the experiments; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Chenxi Li; Fudong Tang; Yangyang Wang; Su Wu; Hui Zhi: Performed the experiments; Wrote the paper.

Fan Zhang: Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Meiyun Wang: Analyzed and interpreted the data; Wrote the paper.

Jiaqiang zhang: Conceived and designed the experiments; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Data availability statement

Data will be made available on request.

Declaration of interest's statement

The authors declare no conflict of interest.

Additional information

No additional information is available for this paper.

Acknowledgements

The authors would like to thank all the reviewers who participated in the review, as well as MJEditor (www.mjeditor.com) for providing English editing services during the preparation of this manuscript.

References

[1] U.b.t.C.o. Standards, et al., Practice guidelines for management of the difficult airway: an updated report by the American society of anesthesiologists task force on management of the difficult airway, Anesthesiology 118 (2) (2013) 251–270.
[2] A. Nørskov, C. Rosenstock, J. Wetterslev, G. Astrup, A. Afshari, L. Lundstrøm, Diagnostic accuracy of anaesthesiologists' prediction of difficult airway management in daily clinical practice: a cohort study of 188 064 patients registered in the Danish Anaesthesia Database, Anaesthesia 70 (3) (2015) 272–281.
[3] O. Langeron, et al., Prediction of difficult mask ventilation, J. Am. Soci. Anesth. 92 (5) (2000) 1229–1236.
[4] S. Kheterpal, et al., Incidence and predictors of difficult and impossible mask ventilation, J. Am. Soci. Anesth. 105 (5) (2006) 885–891.
[5] M.E. Detsky, et al., Will this patient be difficult to intubate?: the rational clinical examination systematic review, JAMA 321 (5) (2019) 493–503.
[6] S. Kheterpal, et al., Incidence, predictors, and outcome of difficult mask ventilation combined with difficult laryngoscopy: a report from the multicenter perioperative outcomes group, Anesthesiology 119 (6) (2013) 1360–1369.
[7] R.M. Levitan, J.W. Heitz, M. Sweeney, R.M. Cooper, The complexities of tracheal intubation with direct laryngoscopy and alternative intubation devices, Ann. Emerg. Med. 57 (3) (2011) 240–247.
[8] T. Cook, N. Woodall, C. Frerk, F.N.A. Project, Major complications of airway management in the UK: results of the fourth national Audit Project of the royal College of anaesthetists and the difficult airway society. Part 1: anaesthesia, Br. J. Anaesth. 106 (5) (2011) 617–631.
[9] T. Cook, S. MacDougall-Davis, Complications and failure of airway management, Br. J. Anaesth. 109 (2012) i68–i85, suppl_1.
[10] D. Edelman, E. Perkins, D. Brewster, Difficult airway management algorithms: a directed review, Anaesthesia 74 (9) (2019) 1175–1185.
[11] T. Heidegger, Management of the difficult airway, N. Engl. J. Med. 384 (19) (2021) 1836–1847.
[12] N. Chrimes, W. Bradley, J. Gatward, A. Weatherall, Human Factors and the 'next Generation'airway Trolley, vol. 74, Wiley Online Library, 2019, pp. 427–433.
[13] A.K. Nørskov, C.V. Rosenstock, L.H. Lundstrøm, Lack of national consensus in preoperative airway assessment, Changes 9 (12) (2016), 13.
[14] Y. Hagiwara, H. Watase, H. Okamoto, T. Goto, K. Hasegawa, J.E.M.N. Investigators, Prospective validation of the modified LEMON criteria to predict difficult intubation in the ED, Am. J. Emerg. Med. 33 (10) (2015) 1492–1496.
[15] A. Nørskov, et al., Effects of using the simplified airway risk index vs usual airway assessment on unanticipated difficult tracheal intubation-a cluster randomized trial with 64,273 participants, Br. J. Addiction: Br. J. Anaesth. 116 (5) (2016) 680–689.
[16] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444.
[17] M.Y. Lu, et al., AI-based pathology predicts origins for cancers of unknown primary, Nature 594 (7861) (2021) 106–110.
[18] A. Esteva, et al., Dermatologist-level classification of skin cancer with deep neural networks, Nature 542 (7639) (2017) 115–118.
[19] V. Gulshan, et al., Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs, JAMA 316 (22) (2016) 2402–2410.
[20] Q. Fu, et al., A deep learning algorithm for detection of oral cavity squamous cell carcinoma from photographic images: a retrospective study, EClinical Med. 27 (2020), 100558.
[21] C.W. Connor, S. Segal, Accurate classification of difficult intubation by computerized facial analysis, Anesth. Analg. 112 (1) (2011) 84–93.
[22] G.L. Cuendet, et al., Facial image analysis for fully automatic prediction of difficult endotracheal intubation, IEEE (Inst. Electr. Electron. Eng.) Trans. Biomed. Eng. 63 (2) (2015) 328–339.
[23] K. Aguilar, G.H. Alférez, C. Aguilar, Detection of difficult airway using deep learning, Mach. Vis. Appl. 31 (1) (2020) 1–11.
[24] T. Hayasaka, K. Kawano, K. Kurihara, H. Suzuki, M. Nakane, K. Kawamae, Creation of an artificial intelligence model for intubation difficulty classification by deep learning (convolutional neural network) using face images: an observational study, J. Int. Care 9 (1) (2021) 1–14.
[25] D.G. Altman, J.M. Bland, Statistics notes: the normal distribution, BMJ 310 (6975) (1995) 298.
[26] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.-C. Chen, Mobilenetv2: inverted residuals and linear bottlenecks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4510–4520.
[27] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, C.A. Raffel, Mixmatch: a holistic approach to semi-supervised learning, Adv. Neural Inf. Process. Syst. 32 (2019).
[28] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: a large-scale hierarchical image database, in: IEEE Conference on Computer Vision and Pattern Recognition, 2009, 2009, pp. 248–255. Ieee.
[29] I. Loshchilov, F. Hutter, Fixing Weight Decay Regularization in Adam, 2018.
[30] G. Zhang, C. Wang, B. Xu, R. Grosse, Three mechanisms of weight decay regularization, arXiv preprint arXiv:1810.12281 (2018).
[31] D.C. Cireşan, U. Meier, L.M. Gambardella, J. Schmidhuber, Deep, big, simple neural nets for handwritten digit recognition, Neural Comput. 22 (12) (2010) 3207–3220.
[32] E.D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, Q.V. Le, Autoaugment: Learning augmentation policies from data, arXiv preprint (2018) *arXiv:1805.09501*.
[33] P.Y. Simard, D. Steinkraus, J.C. Platt, Best practices for convolutional neural networks applied to visual document analysis, Icdar 3 (2003), 2003: Edinburgh.
[34] S. Laine, T. Aila, Temporal ensembling for semi-supervised learning, arXiv preprint (2016) *arXiv:1610.02242*.
[35] M. Sajjadi, M. Javanmardi, T. Tasdizen, Regularization with stochastic transformations and perturbations for deep semi-supervised learning, Adv. Neural Inf. Process. Syst. 29 (2016).
[36] A. Tarvainen, H. Valpola, Mean teachers are better role models: weight-averaged consistency targets improve semi-supervised deep learning results, Adv. Neural Inf. Process. Syst. 30 (2017).
[37] Y. Grandvalet, Y. Bengio, Semi-supervised learning by entropy minimization, Adv. Neural Inf. Process. Syst. 17 (2004).
[38] H. Zhang, M. Cisse, Y.N. Dauphin, D. Lopez-Paz, mixup: beyond empirical risk minimization, arXiv preprint arXiv:1710.09412 (2017).
[39] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4700–4708.
[40] C. Szegedy, et al., Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9.
[41] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2818–2826.
[42] N. Ma, X. Zhang, H.-T. Zheng, J. Sun, Shufflenet v2: practical guidelines for efficient cnn architecture design, in: Proceedings of the European Conference on Computer Vision, ECCV), 2018, pp. 116–131.
[43] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint (2014) *arXiv:1409.1556*.
[44] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[45] T. Miyato, S.-i. Maeda, M. Koyama, S. Ishii, Virtual adversarial training: a regularization method for supervised and semi-supervised learning, IEEE Trans. Pattern Anal. Mach. Intell. 41 (8) (2018) 1979–1993.
[46] D.-H. Lee, Pseudo-label: the simple and efficient semi-supervised learning method for deep neural networks, in: Workshop on challenges in representation learning, ICML, 3, 2013, p. 896, no. 2.
[47] C. Rosenstock, J. Møller, A. Hauberg, Complaints related to respiratory events in anaesthesia and intensive care medicine from 1994 to 1998 in Denmark, Acta Anaesthesiol. Scand. 45 (1) (2001) 53–58.
[48] M. Janssens, G. Hartstein, Management of difficult intubation, Eur. J. Anaesthesiol. 18 (1) (2001) 3–12.
[49] L.M. Ho-Tai, J.H. Devitt, A.G. Noel, M.P. O'Donnell, Gas leak and gastric insufflation during controlled ventilation: face maskversus laryngeal mask airway, Can. J. Anaesth. 45 (3) (1998) 206–211.

[50] J. Weller, A. Merry, B. Robinson, G. Warman, A. Janssen, The impact of trained assistance on error rates in anaesthesia: a simulation-based randomised controlled trial, Anaesthesia 64 (2) (2009) 126–130.

[51] T. Shiga, Z.i. Wajima, T. Inoue, A. Sakamoto, Predicting difficult intubation in apparently normal patients: a meta-analysis of bedside screening test performance, J. Am. Soci. Anesth. 103 (2) (2005) 429–437.

[52] A. Aminian, et al., Association of metabolic surgery with major adverse cardiovascular outcomes in patients with type 2 diabetes and obesity, JAMA 322 (13) (2019) 1271–1282.

[53] G. Campanella, et al., Clinical-grade computational pathology using weakly supervised deep learning on whole slide images, Nat. Med. 25 (8) (2019) 1301–1309.

[54] S. Chilamkurthy, et al., Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study, Lancet 392 (10162) (2018) 2388–2396.

[55] A. Bindra, H. Prabhakar, G.P. Singh, Z. Ali, V. Singhal, Is the modified Mallampati test performed in supine position a reliable predictor of difficult tracheal intubation? J. Anesth. 24 (3) (2010) 482–485.

[56] S.R. Lewis, A.R. Butler, J. Parker, T.M. Cook, A.F. Smith, Videolaryngoscopy versus direct laryngoscopy for adult patients requiring tracheal intubation, Cochrane Database Syst. Rev. 11 (2016).

[57] J. Zhang, W. Jiang, F. Urdaneta, Economic analysis of the use of video laryngoscopy versus direct laryngoscopy in the surgical setting, J. Comp. Eff. Res. 10 (10) (2021) 831–844.