BMC Medical Education

https://doi.org/10.1186/s12909-026-08817-4

Article in Press

Artificial intelligence-based image recognition in bronchoscopy: software development and randomized controlled trial for training evaluation in intensive care residents

Received: 4 May 2025

Accepted: 11 February 2026

Published online: 24 February 2026

Cite this article as: Brunoni B., Zadek F., Pampurini F. *et al.* Artificial intelligence-based image recognition in bronchoscopy: software development and randomized controlled trial for training evaluation in intensive care residents. *BMC Med Educ* (2026). https://doi.org/10.1186/s12909-026-08817-4

Beatrice Brunoni, Francesco Zadek, Federica Pampurini, Marco Vettorello, Francesco Baccoli, Federico Cabitza, Roberto Fumagalli & Thomas Langer

Artificial intelligence-based image recognition in bronchoscopy: software development and randomized controlled trial for training evaluation in intensive care residents

Short title: AI-based versus human-led training for bronchoscopy

Beatrice Brunoni[1], Francesco Zadek[1], Federica Pampurini[2], Marco Vettorello[3], Francesco Baccoli[1], Federico Cabitza[4.5], Roberto Fumagalli[1, 3], Thomas Langer[1, 3]

[1] Department of Medicine and Surgery, University of Milano-Bicocca, Monza, Italy

[2] AI Department, Softlab S.p.A, Milano, Italy

[3] Department of Anesthesia and Intensive Care Medicine, Niguarda Ca' Granda, Milan, Italy

[4] Department of Informatics, Systems and Communication, University of Milano-Bicocca, Milan, Italy

[5] IRCCS Ospedale Galeazzi - Sant'Ambrogio, Milan, Italy

Address for Correspondence: Thomas Langer, MD; Department of Medicine and Surgery, University of Milan-Bicocca, Monza, Italy; Department of Anesthesia and Intensive Care Medicine, Niguarda Ca' Granda, Milan, Italy, Italy. email: Thomas.Langer@unimib.it

Abstract 290/350 words

Background Flexible bronchoscopy is an essential tool for airway management and both diagnostic and therapeutic interventions, particularly in critical care. Accurate identification of tracheobronchial structures is crucial but challenging for less experienced clinicians, often leading to prolonged procedures and increased complication risks. Simulation-based training using virtual reality or manikins has shown promise, and recent studies suggest that artificial intelligence (AI)-based training outperforms self-directed learning. Limited data exist comparing AI-based bronchoscopy training to expert-led instruction. This study aimed to develop and evaluate a custom-made AI-based software for identifying key tracheobronchial structures and assessing its effectiveness as a training tool for anesthesia and intensive care residents.

Methods An AI-based software using YOLOv8 artificial neural networks was developed to recognize key tracheobronchial structures from bronchoscopy videos of a high-fidelity manikin. In a randomized trial, 22 second-year anesthesia residents with limited bronchoscopy experience were assigned to either AI-based unsupervised training (n=11) or traditional human-led training (n=11). Bronchoscopy skills were assessed using the modified Bronchoscopy Skill and Task Assessment Tool (BSTAT) before and after training.

Results The AI model demonstrated high accuracy, with an average precision-recall AUC of 0.98 and a mean average precision of 0.98. Both groups of residents showed significant improvement in their BSTAT scores (from 30±4 to 53±2, $p<0.001$) and reduced procedural time (from 217±44 to 101±23 seconds, $p<0.001$). No significant differences were observed between the AI-based and expert-led training groups.

Conclusion We developed an AI-based software capable of real-time guidance during flexible bronchoscopy. The AI-based training demonstrated comparable efficacy to expert-led instruction, suggesting its potential as a viable tool for unsupervised medical training in flexible bronchoscopy.

Keywords: Bronchoscopy, Artificial intelligence, airway management, simulation training, Bronchi

Background

Flexible bronchoscopy is fundamental in critical care, serving both for complex airway management and as a diagnostic and therapeutic tool [1]. Accurate recognition of the tracheobronchial structures is essential to correctly locate and describe the presence of pathological findings and to perform bronchoalveolar lavage in a selected lung region. Indeed, also the ability and the technique of the operator to navigate within the airways directly affect the time required to complete the procedure [2]. In particular, in critically ill patients, the duration of the procedure is directly associated with the incidence of complications [3], such as hypoxemia, hypercapnia, and hemodynamic instability. Therefore, the importance of achieving a high-level competence through bronchoscopy training [4] is essential for enhancing physicians' skills and improving patient safety [5].

Many clinicians acquire bronchoscopy skills through the apprenticeship model, adhering to the traditional concept of ‘see one, do one, and teach one’ [6]. However, in today’s medical landscape, simulation-based training provides a risk-free platform to learn procedures. For this purpose, training programs based on the use of phantoms, virtual reality simulators [7], and/or electromagnetic navigation systems have been developed and tested on students and junior doctors with promising results [8, 9]. While some reports describe the initial use of artificial intelligence (AI) – based image recognition systems to correctly identify proximal tracheobronchial structures [10, 11], its use for training purposes has only been recently described [12, 13]. These studies demonstrated that the Artificial Neural Network (ANN) has a performance in recognition of airway structures similar to an expert operator and this AI-software can be used as a self-training tool for medical students

learning bronchoscopy [13]. Furthermore, a single study recently compared a *proprietary* AI-based system used for training purposes with an expert-based training for bronchoscopy, showing similar improvements in both groups of trainees [14].

In the present manuscript, we describe in detail the development of a custom-made AI-based software able to correctly identify the principal tracheobronchial structures of a manikin. Moreover, we present its application as a training tool in a randomized controlled trial involving anesthesia and intensive care residents inexperienced in flexible bronchoscopy. We hypothesized that the performance after the AI-based training system would be similar to classical human-led training.

Materials and Methods

This study was performed at the Niguarda Hospital and at the University of Milano Bicocca, both located in Milano, Italy. The randomized controlled trial was retrospectively registered with the ISRCTN registry (ISRCTN63799884).

Development of the AI-based tracheobronchial image recognition software

Manikin, image acquisition, and labeling

An orally intubated manikin with high-fidelity tracheobronchial structures (TruCorp AirSim Advance X model, Lurgan, N. Ireland) was employed. Bronchoscopies were conducted using a flexible bronchoscope (Insighters iS3-C5, Guangdong, Cina) by 10 operators with different degrees of expertise. Acquired bronchoscopy videos were uploaded to the Edge Impulse platform (EdgeImpulse Inc 2023, San Jose California, USA) for segmentation into frames (1 frame every 2 seconds of video) and subsequent labeling. All images deriving from video frames were standardized to a resolution of 480 pixels. Labeling was performed for the following structures: Trachea, Carina, Right Bronchus, Upper Right Lobe Bronchus, Right Truncus Intermedius Bronchus, Right Lower Lobe, Right Middle Lobe, Left Bronchus, Left Secondary Carina, Left Upper Lobe, Left Lower Lobe, Left Upper Division Lobe, Lingula (**Figure 1A**).

Image labeling was performed by two experienced physicians and subsequently verified by a third (illustrated in **e-Figure 1** of the online supplementary material), using the original bronchoscopy videos to guarantee the correct identification of tracheobronchial structures. All images acquired during bronchoscopy were included, even those of the deeper bronchi.

However, no labeling was performed on these structures as they were not the target of image recognition. Of note, each image could have more than one label if it contained more structures, *e.g.*, Right Bronchus and Left Bronchus (**e-Figure 1A**).

Data augmentation was performed to enhance the model's robustness and avoid bias in class predictions. Since the initial dataset had an uneven distribution of images across classes, a manual augmentation step was first applied to increase the number of images in underrepresented classes, ensuring a balanced dataset. Techniques such as rotations (±180°), zoom in/out, translations, and noise addition were employed using Python libraries, with the goal of improving the model's ability to generalize to variations in orientation, scale, and lighting conditions. In addition, automatic augmentation was applied via the Albumentations library, integrated into the YOLOv8 pipeline.

Development of the artificial neural network

The artificial neural network (ANN) for image recognition was developed using YOLOv8 ("You Only Look Once: Unified, Real-Time Object Detection") [15, 16]. The image dataset was divided into a training set (80% of total images) and a validation set (the remaining 20%). Both labeled and non-labeled images were included to improve the model's capacity to distinguish target structures from background anatomy, particularly distal bronchi. **Figure 1B** illustrates the process of developing this AI-image recognition software.

Training

The training process focused on ensuring that the ANN accurately identified and classified key tracheobronchial structures in real-time bronchoscopy scenarios. A customized training loop was developed, exposing the ANN to the dataset across multiple iterations (epochs). During each epoch, the model updated its parameters by progressively improving its performance, using feedback from both correct and incorrect predictions. The training was designed to handle real-time demands, where rapid and accurate recognition is essential for guiding medical procedures.

Throughout the training, the model's performance was continuously monitored using standard evaluation metrics, such as accuracy and precision, ensuring it could reliably differentiate between structures. Regular validation during training allowed for early detection of any performance plateaus, ensuring that the model did not overfit the training data and could maintain its accuracy when applied to new, unseen cases (**e-Figure 2**).

The process was further supported by real-time feedback mechanisms, allowing for rapid adjustments to the training if the model began to show signs of bias or overfitting. This adaptive approach ensured that the ANN could achieve optimal performance, balancing accuracy with the ability to make rapid decisions, a critical requirement for bronchoscopy in clinical settings.

Validation

The model's performance was evaluated through a normalized Confusion Matrix, Precision-Recall curves, and Recall-Confidence curves [10]. The normalized Confusion Matrix compares the AI model's predictions to the true

labels for each image. It captures the model's performance by summarizing the frequency of correct identifications for a given structure and errors, such as failing to detect a structure (false negatives) or falsely identifying one (false positives). Normalization converts raw counts into percentages, facilitating comparison across different categories.

Precision-Recall curves visualize the relationship between Precision, defined as the model's Positive Predictive Value (e.g., when the model predicts that an image contains a certain structure, how often is it correct?), and Recall, or Sensitivity (e.g., of all the images containing the carina, how many did the model correctly identify?). These curves illustrate the trade-off between these two metrics, aiding in balancing prediction accuracy with comprehensive detection [17]. Finally, Recall-Confidence Curves were constructed to show how the model’s ability to correctly identify structures (recall) varies with its confidence in predictions. This analysis was used to fine-tune the confidence threshold, ensuring an appropriate balance between minimizing missed structures and avoiding false positives in subsequent training applications.

To use the developed AI-base image recognition tool for the training of residents we subsequently created a Python application. Identified structures were labeled only if the set confidence threshold was reached. The coverage, defined as the percentage of images assigned by the system a confidence score equal to or above the threshold and thus proposed to users, was calculated for each individual tracheobronchial structure.

Application for the training of residents

Residents in anesthesia and intensive care of the University of Milano-Bicocca without prior exposure to specific bronchoscopy training were enrolled to compare the AI-based to classical human-led training. Consent for the publication of data was obtained from residents. After a 1-hour frontal lecture on bronchoscopy and bronchial anatomy, the baseline bronchoscopy skills of all participants were tested using the Bronchoscopy Skill and Task Assessment Tool (BSTAT) [18]. Specifically, as previously reported by several authors [19–21], we used a modified version, which evaluates only proximal tracheobronchial structures **(Supplementary materials),** on which the AI-based software was specifically trained. Overall, the modified BSTAT evaluates the theoretical knowledge regarding the recognition of proximal bronchial anatomy (28 points total) and a practical component, assessing procedural positioning, airway wall trauma, correct intrabronchial scope position, and access to several tracheobronchial structures (27 points total). Of note, the proportion between the score driven by knowledge and practical component is similar to the original version of the BSTAT. Finally, similarly to the original BSTAT, the time required to complete the examination was recorded.

Participants were thereafter randomized in a 1:1 ratio using sealed envelopes. The first group received classical training performed by an expert bronchoscopy instructor. The second group performed unsupervised training using the AI-based image recognition software. Each resident had 20 minutes of individual training and watched the individual training sessions of the other residents of her/his group. At the end of the training, each resident repeated

the modified BSTAT. Of note, the assessment of the modified BSTAT was always performed by the same person blinded for group allocation.

Both baseline and post-training BSTAT examinations were conducted individually to prevent any learning effect from observation, ensuring that each resident's performance was based solely on their own training experience.

Sample Size

The primary aim of this study was to compare the performance of residents in flexible bronchoscopy after specific training, either AI-based or human-led. Specifically, we hypothesized that the BSTAT scores of residents undergoing AI-based training would be similar to those assigned to human-led training (H0). We considered a difference of 3 points between the two groups as relevant for training purposes. A sample size of 22 residents was calculated considering a significance level (alpha) of 0.05 and a power of 0.9, assuming an absolute difference in score of 3 points between the two training methods using a two-tailed comparison of means with a standard deviation of 2 (effect size 1.5).

Statistical analysis

Data was expressed as mean±standard deviation. The normality of the population was tested via Shapiro-Wilk test. Comparison between the absolute means between the two groups was performed via Welch’s t-test. Comparisons of repeated measurements of continuous variables were performed using pair t-test or Signed Rank Sum Test, as appropriate. Statistical significance was defined as a p-value lower than the significance level ($p < 0.05$). Analyses were performed using Stata statistical software (Stata Statistical Software,

Release 16; StataCorp, College Station, TX, USA). The diagrams were created with Python library seaborn. Images were created using Biorender (Toronto, Canada).

The CONSORT Guidelines statement for simulation-based trials was used to guide the reporting of the results **(Supplementary materials)** [9].

Results

A total of 130 minutes of bronchoscopy were recorded, resulting in 3900 images. Of these images, 82% were labeled, while the remaining 18% were not, as they did not contain anatomical structures targeted by the automated image recognition system. Image augmentation increased the number of labeled structures by an average of 32±21%. For details regarding image augmentation of the single labeled regions, see **e-Figure 3**. Of the resulting 11407 images, 80% (9126) were used to train the ANN, while the remaining 20% (2281) were employed for the subsequent internal validation.

Figure 2 shows the normalized Confusion Matrix, divided for each target structure. The AI-based tool was able to correctly classify the target tracheobronchial structures with very high accuracy, as indicated by the percentages in the diagonal cells, all ranging between 89% and 100%. The principal problem of the system was the incorrect classification of "background" images, *i.e.*, mainly images deriving from distal tracheobronchial structures that were mislabeled as target structures. The Precision-Recall curves for the single target structures and the average Precision-Recall curve are presented in **Figure 3, Panel A**. The curves represent the relationship between Precision (the proportion of true positive predictions among all positive predictions) and Recall (the proportion of correctly identified relevant structures). The area under the curve (AUC) of different items was very high, with values ranging between 0.93 and 0.99. Individual and average Recall-Confidence curves are described in **Figure 3, Panel B**. Based on this analysis, a confidence threshold of ≥85% was chosen for subsequent application of the newly developed software for training purposes. On average, the coverage with a confidence threshold ≥85% was

84±10%. Data on individual structures can be found in the online supplementary material.

Training for residents

The study was performed in February 2024. Twenty-two second-year anesthesia and intensive care residents (aged 28±2 years old, 59% female) were enrolled and randomized in two groups of 11 individuals each.

No difference in age and sex was observed between the two groups (p=0.49 and p=0.14, respectively). All residents had limited experience with flexible bronchoscopy, 55% had never performed a procedure, and 45% had performed less than 5 bronchoscopies during their residency. Expertise was similar in residents assigned to the AI-based and to the traditional human-based training (p=0.45).

In the overall population, the mean modified BSTAT score improved from a baseline of 30±4 to 53±2 points after training (p<0.001). The increase in post-training score was due to a gain of 10±2 points regarding the theoretical component and 13±3 points in practical skills. Furthermore, we observed a marked reduction in procedural time (218±44 vs. 101±23 seconds, p<0.001), *i.e.*, the time needed to complete the assessment.

Pre- and post-training results of the modified BSTAT examination are reported in **Table 1**. Baseline scores were similar among residents of the two groups, and despite the observed significant improvement, no differences were observed at the post-training assessment (p=0.53). Notably, improvements in knowledge, practical skills, and procedural time were similar among groups.

Discussion

We have developed an AI-based image recognition software able to identify with high accuracy the main tracheobronchial structures of a high-fidelity manikin. The software operates in real-time during flexible bronchoscopy, providing guidance during the procedure. Furthermore, we have employed this newly developed tool for simulation-based training in a group of anesthesia and intensive care residents, demonstrating a similar performance compared to traditional human-based training.

Proper medical training is crucial for ensuring patient safety and delivering high-quality medical care, especially in the setting of emergencies and invasive procedures. Indeed, the risks of invasive procedures are significantly influenced by the operator's expertise and procedural time. A clear disparity in complication rates related to bronchoscopy (hypoxemia, worsening hypercapnia, need for deep sedation) has been described between trained and untrained practitioners [2, 3]. Furthermore, the learning curve for new bronchoscopists is very variable, requiring individualized training times to achieve the same level of performance [22]. Given the increasing number of intensive care and anesthesia residents, and the clinical and ethical considerations regarding the classical apprenticeship model, simulation-based bronchoscopy training is gaining popularity [23–25]. However, classical human-based training still requires the availability of expert trainers. In this regard, AI-based image recognition software can significantly assist unsupervised trainees in improving their recognition skills, thus likely expediting procedures.

However, it is essential to thoroughly understand the structural characteristics and principles behind the development of an ANN. This knowledge is crucial not only to ensure the conscious and critical use of such tools, but also to correctly interpret the results they generate [26–28]. Moreover, this understanding aligns with the latest guidelines on the subject, which emphasize the importance of transparency, validation, and the reliability of algorithms used in clinical settings [29]. Of note, most previously published tools[13, 14] focus on real-time feedback during human bronchoscopy using *proprietary* software, without a detailed description of its development and evaluation.

In our manuscript, we present the main features of the developed software and discuss them below. *First*, the AI-software had to accurately identify, in real-time, the most relevant tracheobronchial structures of a dedicated manikin. For this reason, we based our system on YOLOv8 ANN, which is known for its speed in image identification, labeling, and low latency in real-time usage. Another important aspect of our software is its accuracy and performance, as evidenced by the Confusion Matrix and Confidence recall, which show a high capacity to correctly identify the tracheobronchial structures. Different from prior software used for similar purposes[10, 11], our system was also trained with “background” images, *i.e.*, images not containing target structures, such as distal bronchi. As this, it learned that some images do not contain target structures and, therefore, do not require labeling [30]. However, while the accuracy of target structures was extremely high, the software made more errors on distal background images, wrongly labeling them as target structures. This occurred most likely due to anatomical similarities of distal and proximal structures in the employed

manikin. Another aspect that characterizes the newly developed software was its ability to run real-time during bronchoscopy and, thus, to be able to guide the trainee during the procedure.

We used the AI-based system for training second-year anesthesia and intensive care residents and found that their improvement in bronchoscopy skills mirrored that of residents undergoing traditional human-led training. Notably, both groups demonstrated significant increases in their modified BSTAT scores, emphasizing the efficacy of bronchoscopy training. In particular, the increased scores were due to similar improvements in theoretical and practical skills. Additionally, the time required to complete the assessment after training was, on average, 40% shorter than at baseline. Although this difference did not reach statistical significance, it is consistent with previous findings [14] and may suggest that AI-based training allows trainees greater autonomy to explore hand positioning and refine manual skills. Could, therefore, a completely unsupervised AI-based bronchoscopy training be hypothesized? While the software effectively assists trainees in identifying tracheobronchial structures, it does not provide critical feedback on proper bronchoscope handling and hand positioning. Furthermore, an experienced trainer offers invaluable human factors that enhance motivation and likely foster the retention of information.

Two recent studies share similarities with our research, as they explore the role of AI-assisted feedback in improving bronchoscopy performance in a simulated setting [13, 14]. However, there are some differences worth noting. The *first* difference is that previous studies employed the same *proprietary* AI-feedback system for real-time guidance on structured progression and diagnostic completeness. Differently, we developed a custom-made AI system

that can potentially be trained for different bronchoscopes and clinical settings. The *second* aspect is the training modality in the control group. In the study by Cold et al. the AI-guided training was compared with self-directed learning using written instructions, showing that students receiving AI-feedback outperformed their peers. Similarly to our study, Agbontaen et al. directly compared AI-based and expert-led training, demonstrating comparable performance between the two modalities. Of note, both studies employed the proprietary AI-software to assess the performance of trainees, while in our case, rating was expert-based using a modified BSTAT.

Taken together, our and previous studies support that AI-based training can be safely integrated in current training programs. However, we believe that a hybrid training platform—integrating traditional human-based instruction with individual, unsupervised AI-guided sessions—may be the most effective approach for improving clinicians' skills in this invasive procedure [31, 32].

In particular, human-based training is valuable for maintaining motivation and providing feedback on bronchoscope management (e.g., scope handling, operator's position), in fact in some aspect human guidance is better and can help better improvement on long term training [32] . Despite these premises, no significant difference was observed in practical skills after the training.

Future clinical applications

It is conceivable to hypothesize that a similar software could be developed, training the system with human bronchoscopy videos, including both

physiological and pathological images. The bronchoscopy navigation software could certainly assist clinicians in enhancing their accuracy and reducing the time needed to complete the procedure, possibly reducing complications [33]. Indeed, as demonstrated in other medical contexts, such as electrocardiography (ECG) and magnetic resonance imaging (MRI) [34], AI-based image recognition tools can effectively support physicians in clinical decision-making and skill acquisition.

Limitations

Some limitations of our study should be acknowledged. *First*, the accuracy of the software is specific to the used manikin. However, this limitation is mitigated by the potential to retrain the model on other phantoms or even real human images, which would enhance its broader applicability. *Second*, our study included a relatively small sample size and evaluated outcomes only at Kirkpatrick level 2 (i.e., acquisition of knowledge and skills). While the observed consistency across participants suggests generalizability, further studies are needed to evaluate the clinical impact and long-term retention of skills. *Third*, we employed a simplified version of the BSTAT to assess bronchoscopy performance. Although this version was appropriate for evaluating core skills in the acute care setting, it may have introduced a ceiling effect and reduced the ability to detect subtle performance differences between groups. *Finally*, we need to acknowledge that the current version of the software was trained only on the main tracheobronchial structures, while segmental bronchi were not included.

Conclusion

Our AI-based software is able to accurately identify in real-time the main tracheobronchial structures in a high-fidelity manikin, effectively providing procedural guidance. When tested on second-year residents, the AI-based training was not significantly different from the traditional human-led training, suggesting its potential for unsupervised training.

List of abbreviations

AI: artificial intelligence; BSTAT: Bronchoscopy Skill and Task Assessment Tool; AUC: Area under the ROC Curve; ANN: Artificial Neural Network; ECG: electrocardiography; MRI: magnetic resonance imaging; mAP: mean average precision; IoU: intersection-over-union.

Ethical approval and consent to participate.

The need for ethics approval was waived by the Ethics Committee of the University of Milano-Bicocca. The study was conducted in accordance with the guidelines outlined in the Declaration of Helsinki. Consent for participation in the study was obtained from all Intensive Care and Anesthesia residents.

Clinical trial number: ISRCTN63799884.

Consent for publication

Not applicable

Availability of data and materials statement

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Conflicts of interest: none

Competing interest: The authors have no conflicts of interest to declare.

Funding Sources: No funding.

Author contribution statement

Concept and study design: TL, BB, RF; AI-software development: BB, FP, MV; Training and data recruitment: BB, FP, FB, MV; Data analysis and interpretation: BB, FZ, FP;FC Drafting of article: BB, TL; Designing of figures: BB, FZ, FP; Critical review, editing and approval of article: all authors. All authors confirm that this article, including related data and figures, has not been previously reported or published elsewhere.

Acknowledgement The authors are indebted to Alessandro and Anna Meazza for their kind donation, which made this work possible, and to Valentina Palladio for her contributions during her medical student training

Reference

1. Patolia S, Farhat R, Subramaniyam R (2021) Bronchoscopy in intubated and non-intubated intensive care unit patients with respiratory failure. J Thorac Dis 13:5125–5134. https://doi.org/10.21037/jtd-19-3709
2. Ouellette DR (2006) The Safety of Bronchoscopy in a Pulmonary Fellowship Program. Chest 130:1185–1190. https://doi.org/10.1378/chest.130.4.1185
3. STATHER DR, MACEACHERN P, CHEE A, et al (2013) Trainee impact on advanced diagnostic bronchoscopy: An analysis of 607 consecutive procedures in an interventional pulmonary practice. Respirology 18:179–184. https://doi.org/10.1111/j.1440-1843.2012.02270.x
4. Shore D, Patel P, Ahmad D (2024) Assessment of bronchoscopy training and competency: a narrative review. AME Med J 9:16–16. https://doi.org/10.21037/amj-23-147
5. Follmann A, Pereira CB, Knauel J, et al (2019) Evaluation of a bronchoscopy guidance system for bronchoscopy training, a randomized controlled trial. BMC Med Educ 19:430. https://doi.org/10.1186/s12909-019-1824-3
6. Ayub SM (2022) “See one, do one, teach one”: Balancing patient care and surgical training in an emergency trauma department. J Glob Health 12:03051. https://doi.org/10.7189/jogh.12.03051
7. Andersen AG, Rahmoui L, Dalsgaard T-S, et al (2023) Preparing for Reality: A Randomized Trial on Immersive Virtual Reality for Bronchoscopy Training. Respiration 102:316–323. https://doi.org/10.1159/000528319

8. Follmann A, Pereira CB, Knauel J, et al (2019) Evaluation of a bronchoscopy guidance system for bronchoscopy training, a randomized controlled trial. BMC Med Educ 19:430. https://doi.org/10.1186/s12909-019-1824-3
9. Cheng A, Kessler D, Mackinnon R, et al (2016) Reporting guidelines for health care simulation research: Extensions to the CONSORT and STROBE statements. BMJ Simul Technol Enhanc Learn 2:51–60. https://doi.org/10.1136/bmjstel-2016-000124
10. Yoo JY, Kang SY, Park JS, et al (2021) Deep learning for anatomical interpretation of video bronchoscopy images. Sci Rep 11:23765. https://doi.org/10.1038/s41598-021-03219-6
11. Li Y, Zheng X, Xie F, et al (2022) Development and validation of the artificial intelligence (AI)-based diagnostic model for bronchial lumen identification. Transl Lung Cancer Res 11:2261–2274. https://doi.org/10.21037/tlcr-22-761
12. Cold KM, Agbontaen K, Nielsen AO, et al (2024) Artificial intelligence for automatic and objective assessment of competencies in flexible bronchoscopy. J Thorac Dis 16:5718–5726. https://doi.org/10.21037/jtd-24-841
13. Cold KM, Xie S, Nielsen AO, et al (2024) Artificial Intelligence Improves Novices' Bronchoscopy Performance. Chest 165:405–413. https://doi.org/10.1016/j.chest.2023.08.015
14. Agbontaen KO, Cold KM, Woods D, et al (2025) Artificial Intelligence-Guided Bronchoscopy is Superior to Human Expert Instruction for the Performance of Critical-Care Physicians: A Randomized Controlled Trial. Crit Care Med 53:e1105–e1115. https://doi.org/10.1097/CCM.0000000000006629
15. Redmon J, Divvala S, Girshick R, Farhadi A (2016) You Only Look Once: Unified, Real-Time Object Detection. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp 779–788
16. https://yolov8.com
17. Saito T, Rehmsmeier M (2015) The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on

Imbalanced Datasets. PLoS One 10:e0118432. https://doi.org/10.1371/journal.pone.0118432

18. Davoudi M, Osann K, Colt HG (2008) Validation of Two Instruments to Assess Technical Bronchoscopic Skill Using Virtual Reality Simulation. Respiration 76:92–101. https://doi.org/10.1159/000126493
19. Schertel A, Geiser T, Hautz WE (2021) Man or machine? Impact of tutor-guided versus simulator-guided short-time bronchoscopy training on students learning outcomes. BMC Med Educ 21:123. https://doi.org/10.1186/s12909-021-02526-w
20. Siow WT, Tan G, Loo C, et al (2021) Impact of structured curriculum with simulation on bronchoscopy. Respirology 26:597–603. https://doi.org/10.1111/resp.14054
21. Colt HG, Davoudi M, Murgu S, Zamanian Rohani N (2011) Measuring learning gain during a one-day introductory bronchoscopy course. Surg Endosc 25:207–216. https://doi.org/10.1007/s00464-010-1161-4
22. Wahidi MM, Silvestri GA, Coakley RD, et al (2010) A Prospective Multicenter Study of Competency Metrics and Educational Interventions in the Learning of Bronchoscopy Among New Pulmonary Fellows. Chest 137:1040–1049. https://doi.org/10.1378/chest.09-1234
23. Silvestri GA (2008) The Evolution of Bronchoscopy Training. Respiration 76:19–20. https://doi.org/10.1159/000127579
24. Kronborg SH, Karbing DS, Arshad A, Lundgaard AC (2024) Four different models for simulation-based training of bronchoscopic procedures. BMC Pulm Med 24:. https://doi.org/10.1186/s12890-024-02846-9
25. Cold KM, Konge L, Clementsen PF, Nayahangan LJ (2019) Simulation-Based Mastery Learning of Flexible Bronchoscopy: Deciding Factors for Completion. Respiration 97:160–167. https://doi.org/10.1159/000493431
26. Barua I, Wieszczy P, Kudo S, et al (2022) Real-Time Artificial Intelligence–Based Optical Diagnosis of Neoplastic Polyps during Colonoscopy. NEJM Evidence 1:. https://doi.org/10.1056/EVIDoa2200003
27. Patrini I, Ruperti M, Moccia S, et al (2020) Transfer learning for informative-frame selection in laryngoscopic videos through learned

features. Med Biol Eng Comput 58:1225–1238. https://doi.org/10.1007/s11517-020-02127-7

28. Li J, Jiang P, An Q, et al (2024) Medical image identification methods: A review. Comput Biol Med 169:107777. https://doi.org/10.1016/j.compbiomed.2023.107777
29. Antonelli G, Libanio D, De Groof AJ, et al (2025) QUAIDE - Quality assessment of AI preclinical studies in diagnostic endoscopy. Gut 74:153–161. https://doi.org/10.1136/gutjnl-2024-332820
30. Cold KM, Vamadevan A, Laursen CB, et al (2025) Artificial intelligence in bronchoscopy: a systematic review. European Respiratory Review 34:240274. https://doi.org/10.1183/16000617.0274-2024
31. Bhutoria A (2022) Personalized education and Artificial Intelligence in the United States, China, and India: A systematic review using a Human-In-The-Loop model. Computers and Education: Artificial Intelligence 3:100068. https://doi.org/10.1016/j.caeai.2022.100068
32. Su P, He H, Liang Y, et al (2025) Feedback from Human Instructors Is Superior to Guidance by a Virtual Reality Simulator When Learning Flexible Bronchoscopy: A Randomized Controlled Trial. Respiration 104:701–707. https://doi.org/10.1159/000546827
33. Calisto FM, Nunes N, Nascimento JC (2022) Modeling adoption of intelligent agents in medical imaging. Int J Hum Comput Stud 168:102922. https://doi.org/10.1016/j.ijhcs.2022.102922
34. Cabitza F, Campagner A, Ronzio L, et al (2023) Rams, hounds and white boxes: Investigating human–AI collaboration protocols in medical diagnosis. Artif Intell Med 138:102506. https://doi.org/10.1016/j.artmed.2023.102506

Table 1. **Comparison of traditional and AI-based training in a group of 22-second-year anesthesia and intensive care residents**. The residents were divided into two groups: traditional training and Artificial intelligence (AI)- -based training. The table presents the pre-training and post-training scores for overall performance, theoretical knowledge, practical skills and procedural time, evaluated using a modified Bronchoscopy Skills and Task Assessment Tool (BSTAT-modified). Additionally, the difference between post and pre-training scores was assessed for each group and each score. Results are reported as means□ standard deviations. The p-values were calculated

		Traditional training (N=11)	AI-based training (N=11)	*P-values*	Difference of mean	95% CI
Pre-training	**Score (points)**	30 □ 5	29 □ 4	0.64	-0.9	-4.9; 3.1
	Theoretical (points)	18 □ 3	18 □ 2	1.00	0.0	-2.3; 2.3
	Practical skills (points)	13 □ 3	12 □ 2	0.44	-0.9	-3.3; 1.5
	Procedural time (seconds)	215 □ 32	222 □ 56	0.73	7	-34; 48
Post-training	**Score (points)**	52 □ 2	53 □ 2	0.53	0.5	-1.0; 1.9
	Theoretical (points)	28 □ 1	28 □ 1	1.00	0.0	-0.3; 0.4
	Practical skills (points)	25 □ 2	25 □ 2	0.53	0.5	-1.0; 1.9
	Procedural time (seconds)	108 □ 24	95 □ 21	0.19	-13	-33; 7
Difference Post-Pre training	**Score (points)**	22 □ 5	23 □ 4	0.45	1.4	-2.3; 5.1
	Theoretical (points)	10 □ 3	10 □ 2	1.00	0.0	-2.3; 2.3
	Practical skills (points)	12 □ 3	13 □ 3	0.27	1.4	-1.2; 3.9
	Procedural time (seconds)	-107 □ 29	-127 □ 58	0.32	20	-61; 21

using a Welch's t-test. 95% confidence interval (CI) for the mean differences are provided.

The scores are presented as Mean modified-BSTAT scores ± SD (standard deviation).

Figure 1. **Methods of development of AI-based image recognition software for bronchoscopy.**

Panel A. Labeling of tracheobronchial structures in TruCorp AirSim Advance X manikin.

Panel B. Overview of the development process for AI-based bronchoscopy image recognition software. 1. Frame acquisition: Bronchoscopy video is captured from the manikin (1a) and individual frames are extracted (2a). 2. Labeling and augmentation: the extracted frames are manually labeled to identify relevant structures (2a), followed by image augmentation to increase the dataset size and variability (2b). 3. Neural network training: the labeled and augmented dataset is used to train a neural network. 4. AI-based software validation: the validation set was used to study the accuracy of the software.

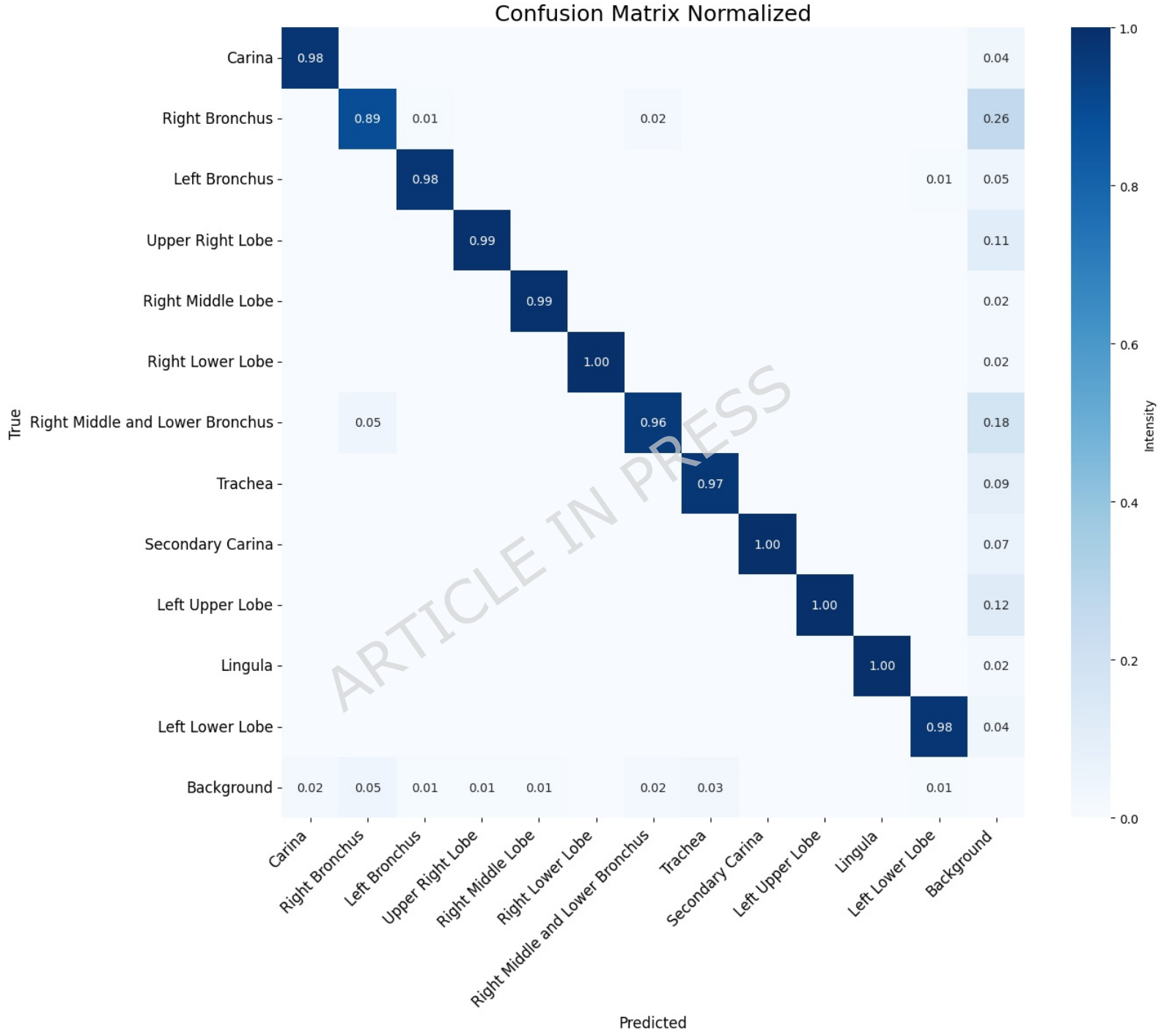
Figure 2. Normalized Confusion Matrix for Image Recognition of main bronchial structures. The matrix illustrates the performance of the image recognition model on the validation set. The y-axis represents the AI predictions, while the X-axis is the true label. Diagonal values indicate the proportion of correct classification, whereas off-diagonal values reflect the misclassifications for each tracheobronchial structure. The blue gradient ranges from light blue 0 to dark blue 1, indicating perfect classification.

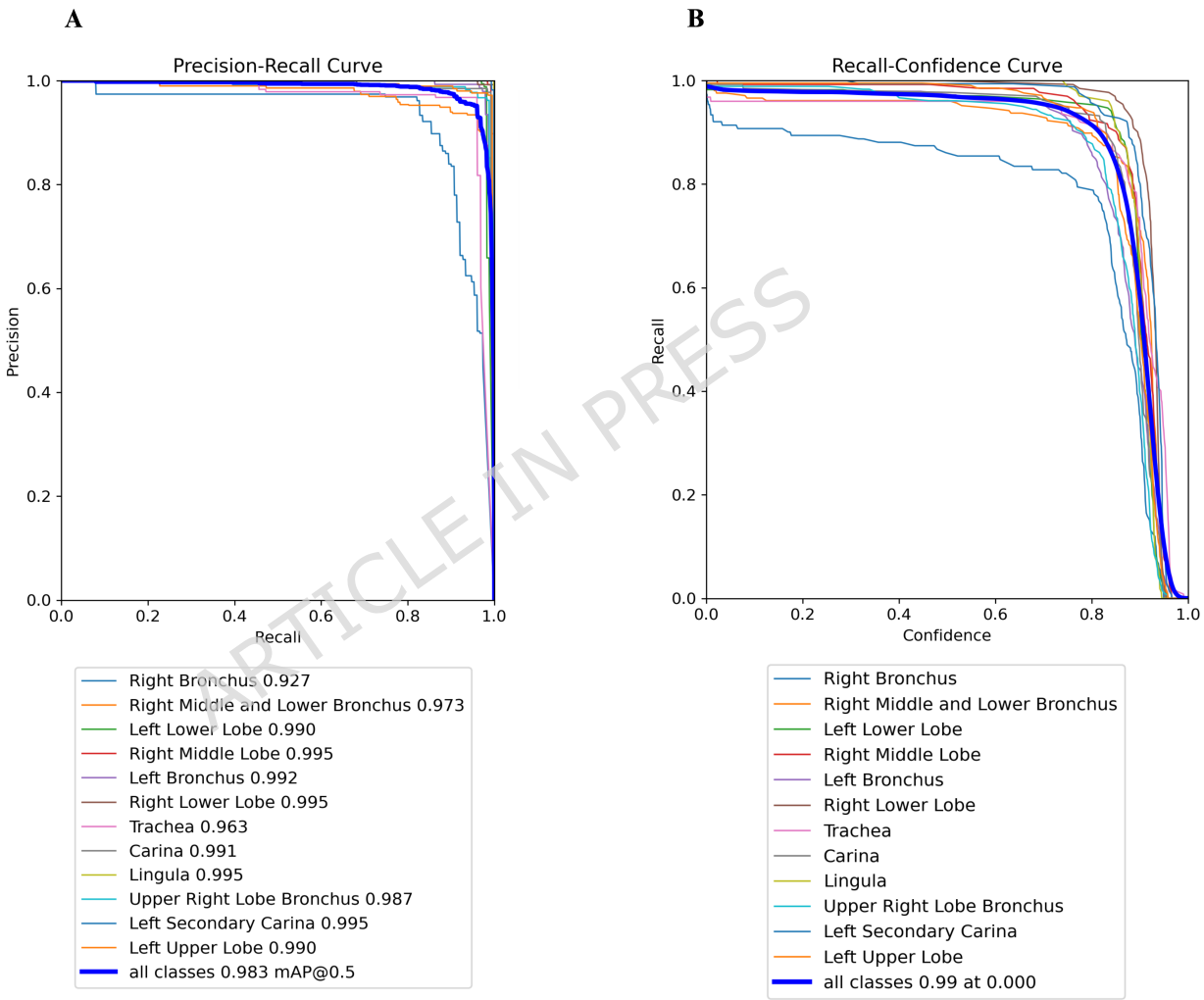
Figure 3. Performance evaluation curves for the image recognition model.

Panel A: Precision-Recall curves for each tracheobronchial structure and all classes.

The curves represent the relationship between **Precision** (the proportion of true positive predictions among all positive predictions) and **Recall** (the proportion of correctly identified relevant structures). Each line in the graph corresponds to a specific tracheobronchial structure, as indicated in the legend. The average performance across all classes is represented by the **mAP@0.5** (mean average precision at an intersection-over-union (IoU) threshold of 0.5), which in this case is 0.983. The **mAP@0.5** summarizes the model's ability to identify structures accurately across different IoU thresholds, where a higher value represents better overall performance.

Panel B. Recall-Confidence curves for each tracheobronchial structure and all classes. The curve illustrates how the model's ability to correctly identify structures (**Recall**) changes with varying levels of **Confidence** in its predictions. Each line corresponds to a specific structure, as shown in the legend. For instance, the **Left Upper Lobe** maintains a recall close to 1.0 even at high confidence levels, while the **Right Bronchus** shows a more gradual decline. The blue line shows the overall performance across all classes, with a recall of 0.99 at a confidence threshold of 0. The plot demonstrates how the model's recall decreases as the confidence threshold increases, helping to assess the balance between recall and confidence for each structure.





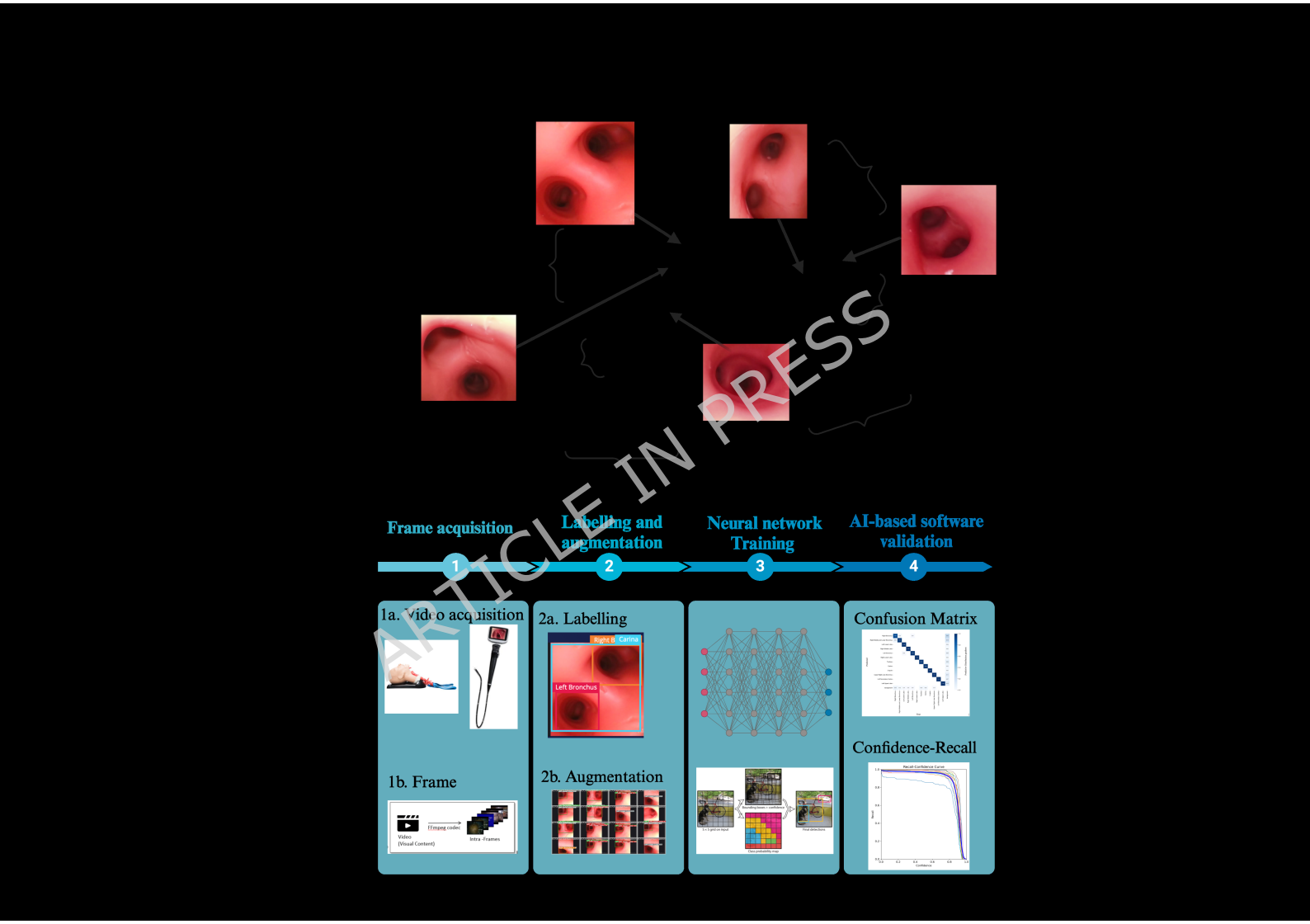


Table 1.

		Traditional training (N=11)	AI-based training (N=11)	*P-values*	Difference of mean	95% CI
Pre-training	**Score (points)**	30 ± 5	29 ± 4	0.64	-0.9	-4.9; 3.1
	Theoretical (points)	18 ± 3	18 ± 2	1.00	0.0	-2.3; 2.3
	Practical skills (points)	13 ± 3	12 ± 2	0.44	-0.9	-3.3; 1.5
	Procedural time (seconds)	215 ± 32	222 ± 56	0.73	7	-34; 48
Post-training	**Score (points)**	52 ± 2	53 ± 2	0.53	0.5	-1.0; 1.9
	Theoretical (points)	28 ± 1	28 ± 1	1.00	0.0	-0.3; 0.4
	Practical skills (points)	25 ± 2	25 ± 2	0.53	0.5	-1.0; 1.9
	Procedural time (seconds)	108 ± 24	95 ± 21	0.19	-13	-33; 7
Difference Post-Pre training	**Score (points)**	22 ± 5	23 ± 4	0.45	1.4	-2.3; 5.1
	Theoretical (points)	10 ± 3	10 ± 2	1.00	0.0	-2.3; 2.3
	Practical skills (points)	12 ± 3	13 ± 3	0.27	1.4	-1.2; 3.9
	Procedural time (seconds)	-107 ± 29	-127 ± 58	0.32	20	-61; 21

The scores are presented as Mean modified-BSTAT scores ± SD (standard deviation).