Brazilian Journal of Anesthesiology 2022;**72(5)**:622–628

SBA
Sociedade Brasileira
de Anestesiologia

Brazilian Journal of
ANESTHESIOLOGY

ORIGINAL INVESTIGATION

Predictive model for difficult laryngoscopy using machine learning: retrospective cohort study

Jong Ho Kim [a,b], **Jun Woo Choi** [a], **Young Suk Kwon** [a,b,*], **Seong Sik Kang** [c]

[a] *Chuncheon Sacred Heart Hospital, Department of Anesthesiology and Pain Medicine, Chuncheon, South Korea*
[b] *Hallym University, Institute of New Frontier Research Team, Chuncheon, South Korea*
[c] *Kangwon National University, College of Medicine, Department of Anesthesiology and Pain Medicine, Chuncheon, South Korea*

Received 22 July 2020; accepted 20 June 2021
Available online 9 July 2021

KEYWORDS
Intratracheal intubation;
Laryngoscopes;
Machine learning

Abstract

Background: Both predictions and predictors of difficult laryngoscopy are controversial. Machine learning is an excellent alternative method for predicting difficult laryngoscopy. This study aimed to develop and validate practical predictive models for difficult laryngoscopy through machine learning.
Methods: Variables for the prediction of difficult laryngoscopy included age, Mallampati grade, body mass index, sternomental distance, and neck circumference. Difficult laryngoscopy was defined as grade 3 and 4 by the Cormack-Lehane classification. Pre-anesthesia and anesthesia data of 616 patients who had undergone anesthesia at a single center were included. The dataset was divided into a base training set (n = 492) and a base test set (n = 124), with equal distribution of difficult laryngoscopy. Training data sets were trained with six algorithms (multilayer perceptron, logistic regression, supportive vector machine, random forest, extreme gradient boosting, and light gradient boosting machine), and cross-validated. The model with the highest area under the receiver operating characteristic curve (AUROC) was chosen as the final model, which was validated with the test set.
Results: The results of cross-validation were best using the light gradient boosting machine algorithm with Mallampati score x age and sternomental distance as predictive model parameters. The predicted AUROC for the difficult laryngoscopy class was 0.71 (95% confidence interval, 0.59–0.83; $p = 0.014$), and the recall (sensitivity) was 0.85.
Conclusion: Predicting difficult laryngoscopy is possible with three parameters. Severe damage resulting from failure to predict difficult laryngoscopy with high recall is small with the reported model. The model's performance can be further enhanced by additional data training.

[*] Corresponding author.
E-mail: gettys@hallym.or.kr (Y.S. Kwon).

https://doi.org/10.1016/j.bjane.2021.06.016

Introduction

Difficult airway (DA) is a clinical situation in which a traditionally trained anesthesiologist experiences difficulty in ventilating the facial mask of the upper airway, intubation, or both.[1] The incidence of difficult laryngoscopy, difficult intubation, and failed intubation was 12.3%, 9%, and 0.47%, respectively.[2] Although difficult airways are uncommon in anesthesia or resuscitation, improper airway management can result in fatal outcomes. Therefore, predicting DA in patients requiring airway management becomes crucial. DA includes difficult laryngoscopy (DL), in which case visualizing any portion of the vocal cords is not possible.[1] Many methods and guidelines for evaluating and predicting DL exist[1,3–8]; however, there is still debate on the best methods and anatomical landmarks to predict DL.[9]

Machine learning (ML) is a category of algorithms that allows accurate software applications predict results. Machine learning is increasingly used for medical diagnose and disease predictions. The basic premise of machine learning involves writing an algorithm that receives input data, uses statistical analysis to predict an output, and updates it as new data becomes available. ML has the power to quickly synthesize and analyze complex multivariable combinations. Thanks to ML, predicting difficult intubation is also changing.[10]

There are several predictive models for DL or DA. However, the use of multiple predictors reduces practicality,[10–14] and most models lack sufficient validation.[10,11] This study aimed to develop and validate practical predictive models for DL through ML using only a few predictors.

Methods

This retrospective cohort study was approved by the Institutional Review Board/Ethics Committee of the Chuncheon Sacred Heart Hospital, Hallym University (IRB No. 2019-07-012-001). The requirement for written informed consent was waived by the ethics committee. The study was conducted retrospectively, and the diagnosis of DL was considered predictors and evaluated by comparing it with the diagnosis of anesthesia record. Data of patients who had undergone surgery under anesthesia at Hallym University Chuncheon Sacred Heart Hospital between January 18, 2019 and November 6, 2019 were collected from preanesthesia and anesthesia records. Exclusion criteria were as follows: < 18 years old, regional anesthesia, major external facial or neck abnormalities, laryngeal abnormalities or tumors, laryngeal mask or mask ventilation, use of video laryngoscopes or fiberoptic scope for tracheal intubation, intubation performed by a resident with < 2 years of anesthesiology experience, insufficient records, and endotracheal intubated states before anesthesia. Of the 3676 selected patients, 3060 were excluded, and 616 were finally included in the study.

Parameters for prediction of difficult laryngoscopy

The parameters for prediction of DL included age, body mass index (BMI), Mallampati grade, sternomental distance (SMD), and neck circumference (NC). The Mallampati score classifies the ease of endotracheal intubations as Class I glottis – fully exposed glottis including the anterior and posterior commissures; Class 2 glottis – partly exposed glottis and not visualized anterior commissure; Class 3 glottis – glottis not exposed and only corniculate cartilages visualized; and Class 4 glottis – corniculate cartilages not exposed. SMD was defined as the straight distance between the upper border of the manubrium sterni and the bony point of the mentum after fully extending the neck with the mouth closed.[8] Neck circumference was measured in cm at the level of the thyroid cartilage.[7] All missing values were removed.

Intubation and laryngoscopy grade

The routine procedures for tracheal intubation are standardized at our hospital. If the vocal cord was not visualized at first attempt, the anesthesiologist would try the laryngoscope manipulation again. Then, depending on the situation, a laryngoscope manipulation was attempted by another anesthesiologist or a video laryngoscope was attempted. Difficult laryngoscopy was defined when all direct laryngoscopy attempts did not visualize the vocal cords. In the anesthesia record, if multiple attempts are made, the number of attempts is recorded and the final Cormack-Lehane grade is recorded. Standard Macintosh metallic single-use disposable laryngoscope blades (INT; Intubrite Llc, Vista, CA, USA) were employed. In total, seven attending anesthesiologists and three resident anesthesiologists classified direct laryngoscopy views in accordance with the Cormack-Lehane grades as Grade 1 = most of the glottic opening is visible; Grade 2 = only the posterior portion of the glottis or only arytenoid cartilages are visible; Grade 3 = only the epiglottis but no portion of the glottis is visible; Grade 4 = neither the glottis nor the epiglottis is visible. Cormack-Lehane 3 and 4 indicated DL and were combined into the difficult class. Cormack-Lehane 1 and 2 were combined into the non-difficult class.

Imbalance learning

The five prediction parameters mentioned were included as variables. A dataset was created with the variables of 616 patients. However, the difficult class (n = 64) included less patients than the non-difficult class (n = 552), resulting in imbalanced data. To solve this problem, we combined over- and under-sampling by the synthetic minority oversampling technique (SMOTE).[15,16]

Variables and algorithm selection

The dataset was randomly divided into a base training set (80%, 492) and a base test set (20%, 124) with equal distribution of difficult class patient data. Through logistic regression of the base training set, we found variables with significant odds ratios (p < 0.05) for DL occurrence and created several training datasets with different variable combinations. We used the base test set to create test sets corresponding to the variables in each training

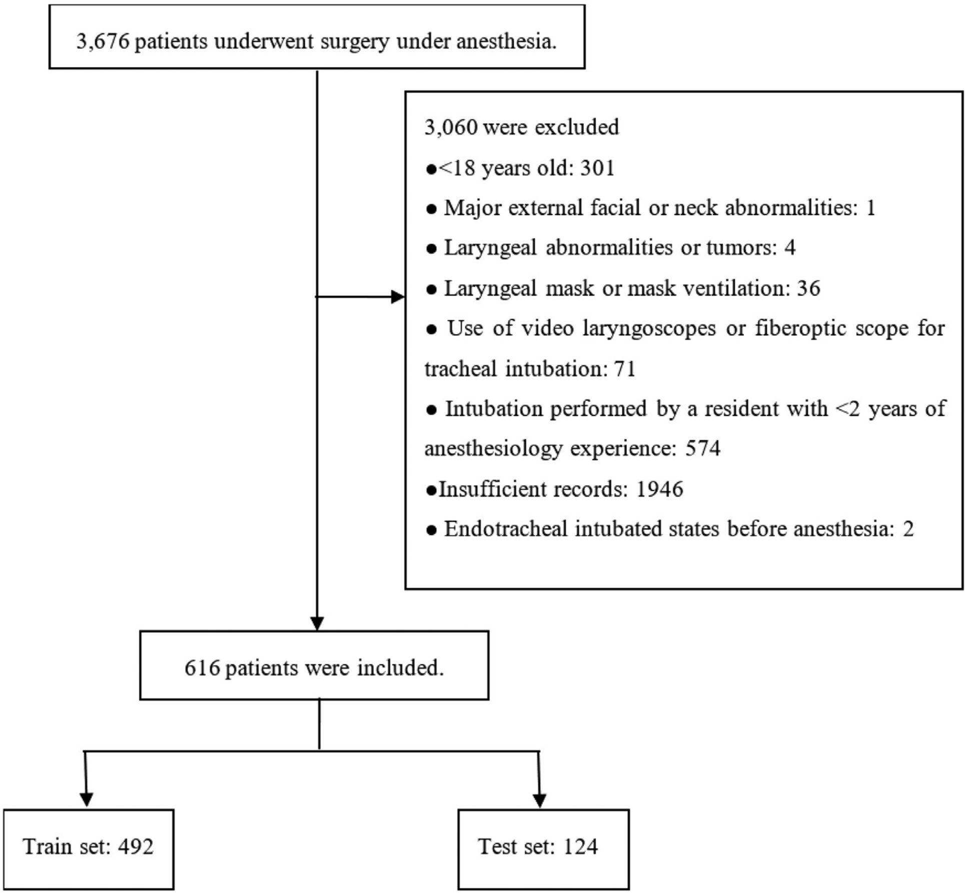
3,676 patients underwent surgery under anesthesia.
3,060 were excluded
●<18 years old: 301
● Major external facial or neck abnormalities: 1
● Laryngeal abnormalities or tumors: 4
● Laryngeal mask or mask ventilation: 36
● Use of video laryngoscopes or fiberoptic scope for tracheal intubation: 71
● Intubation performed by a resident with <2 years of anesthesiology experience: 574
●Insufficient records: 1946
● Endotracheal intubated states before anesthesia: 2
616 patients were included.
Train set: 492
Test set: 124

Figure 1 Flow diagram.

set. Because the purpose of this study is to develop a predictive model with a small number of variables and good performance, a model trained with the basic training set is needed to compare with a model with fewer variables than the basic training set. Also, in addition, a corresponding test set is required to evaluate each training set, including the basic training set. Each training set was normalized by min−max scaling after applying SMOTE, while each test set was normalized by the min−max scaling of the training set. All training sets were trained with six algorithms including multilayer perceptron, logistic regression, supportive vector machines, balanced random forest, extreme gradient boosting, and light gradient boosting machines (LGBM). The developed prediction models applied training sets to algorithms, after which a 10-fold cross-validation was performed and the area under curve of the receiver operating characteristic curve (AUROC) calculated. The results were evaluated by the mean AUROC, and the best model chosen as the final one.

Model evaluation and statistics

The final model was completed after tuning the hyperparameters of the selected model. The performance of the complete model was evaluated through a separate test set, and its performance evaluated by AUROC. Additionally, precision (positive predictive value, ratio of patients with true DL among patients predicted as DL) and recall (sensitivity, ratio of patients predicted as DL among true DL) were calculated. Underdiagnosis is DL, but it is predicted and diagnosed as NDL. If DL leads to airway management failure, it can have serious consequences for the patient. Conversely, overdiagnosis is predicting and diagnosing NDL as DL, and because the practitioner recognize that airway management may be difficult and prepare for it, investment in time, manpower, and equipment due to overdiagnosis may be wasted.

All data were processed and analyzed by Anaconda (Python version 3.7; https://www.anaconda.com). Data set variables were analyzed by SPSS (version 26.0, IBM). Continuous variables were compared with the *t*-test or Mann-Whitney test and categorical variables with the chi-squared test. The odds ratio of the variables was determined by logistic regression (method: enter).

Results

Figure 1 shows enrolment, progress and allocation of study. Table 1 shows the patient's DL prediction parameters in the base training and base test dataset. Table 2 shows the odds ratio of DL variables in the base training set. Age, Mallampati grade, and SMD showed statistical significance. A separate training set was created from the base training set using the three statistically significant variables. In addition, by combining the variables, new potential variables and training

Table 1 Prediction variables of difficult laryngoscopy in base train and base test datasets.

	Base train dataset (n = 492)			Base test sets (n = 124)		
	NDL (n = 441)	DL (n = 51)	*p*-value	NDL (n = 111)	DL (n = 13)	*p*-value
Age (years, mean ± SD)	52.7 ± 16.4	57.6 ± 14.9	0.044	54.2 ± 16.9	68.5 ± 10.8	0.004
Mallampati grade (number, %)			0.008			0.057
Grade 1	97 (22.0)	7 (13.7)		30 (27.0)	2 (15.4)	
Grade 2	167 (37.9)	13 (25.5)		46 (41.4)	3 (23.1)	
Grade 3	115 (26.1)	19 (37.3)		23 (20.7)	5 (38.5)	
Grade 4	62 (14.1)	12 (23.5)		12 (10.8)	3 (23.1)	
SMD (cm, years, mean ± SD)	17.5 ± 2.3	16.6 ± 1.9	0.006	17.3 ± 2.0	16.4 ± 1.5	0.103
NC (cm, years, mean ± SD)	37.1 ± 3.6	37.3 ± 3.3	0.628	37.3 ± 5.1	37.8 ± 3.1	0.717
BMI	25.5 ± 3.9	25.3 ± 4.0	0.714	25.6 ± 4.0	25.6 ± 4.6	0.723

NDL, not difficult laryngoscopy class; DL, difficult laryngoscopy class; SMD, sternomental distance; NC, neck circumference; BMI, body mass index; SD standard deviation.

Table 2 Odds ratio of variables in the base train set.

		Odds ratio (95% CI)	*p*-value
Age		1.02 (1.00–1.04)	0.046
Mallampati grade			0.049
	Grade 1	reference	
	Grade 2	1.08 (0.42–2.80)	0.876
	Grade 3	2.29 (0.92–5.68)	0.074
	Grade 4	2.68 (1.00–7.18)	0.050
SMD		0.81 (0.70–0.94)	0.005
NC		1.02 (0.94–1.10)	0.627
BMI		0.99 (0.91–1.06)	0.713

SMD, Sternomental distance; NC, neck circumference; BMI, body mass index.

Table 3 Ten-fold cross validation AUROC after applying each training set to each algorithm.

	Base train set	Train set 1	Train set 2	Train set 3	Train set 4
MLP (mean ± SD)	0.66 ± 0.03	0.60 ± 0.02	0.62 ± 0.04	0.60 ± 0.03	0.62 ± 0.05
LR (mean ± SD)	0.69 ± 0.1	0.63 ± 0.08	0.75 ± 0.09	0.71 ± 0.15	0.67 ± 0.19
SVM (mean ± SD)	0.68 ± 0.09	0.63 ± 0.08	0.75 ± 0.1	0.72 ± 0.18	0.68 ± 0.19
BRF (mean ± SD)	0.92 ± 0.08	0.90 ± 0.09	0.98 ± 0.03	0.98 ± 0.03	0.98 ± 0.04
XGB (mean ± SD)	0.86 ± 0.1	0.84 ± 0.13	0.99 ± 0.03	0.95 ± 0.05	0.98 ± 0.04
LGBM (mean ± SD)	0.94 ± 0.08	0.92 ± 0.12	0.99 ± 0.02[a]	0.96 ± 0.06	0.97 ± 0.03

Base train set: Mallampati grade, age, sternomental distance, body mass index, neck circumference.
Train set 1: Mallampati grade, age, sternomental distance.
Train set 2: Mallampati grade x age, sternomental distance.
Train set 3: Mallampati grade x sternomental distance, age.
Train set 4: Mallampati grade, sternomental distance x age.
AUROC, area under the receiver operating characteristic curve; MLP, multilayer perceptron; LR, logistic regression; SVM, supportive vector machine; BRF, balanced random forest; XGB, extreme gradient boosting; LGBM, light gradient boosting machine.

[a] The model applying training set 2 to LGBM showed the best performance in cross-validation.

sets based on the new combination variables were created. The variables included in the data set were the following:

- Base train set & Base test set: Mallampati grade, age, SMD, BMI, NC
- Train set 1 & Test set 1: Mallampati grade, age, SMD
- Train set 2 & Test set 2: Mallampati grade x age, SMD
- Train set 3 & Test set 3: Mallampati grade x SMD, age
- Train set 4 & Test set 4: Mallampati grade, SMD x age

The results of cross-validation after applying each train set to each algorithm are shown in Table 3. Model performance was best when Mallampati grade x age and SMD were applied to LGBM in the cross-validation. Therefore, we selected the model applying Mallampati grade x age, and SMD variables to LGBM as the prediction model for DL. The final model was completed after hyperparameter tuning. The test set was applied to the final model to evaluate the final model's performance, which showed an AUROC of 0.71 (95% confidence interval, 0.59–0.83; $p = 0.014$) (Fig. 2). The

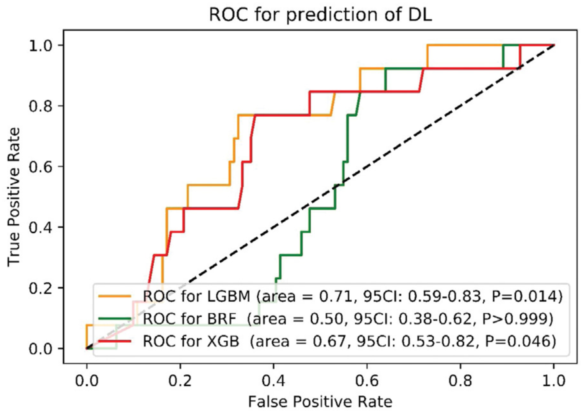

Figure 2 AUROC for difficult laryngoscopy prediction model algorithms: light gradient boosting machine (LGBM), extreme gradient boosting (XGB), balanced random forest (BRF). Predictors: Mallampati grade x age, sternomental distance (SMD). AUROC, area under the receiver operating characteristic curve; DL, difficult laryngoscopy; NDL, no difficult laryngoscopy; 95CI, 95% confidence interval.

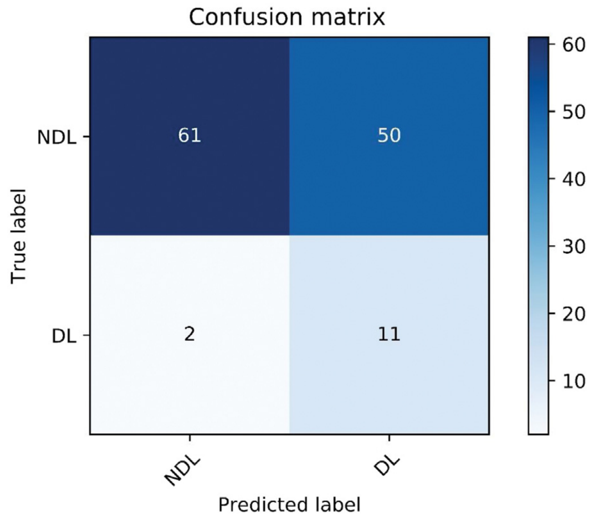

Figure 3 Confusion matrix of difficult laryngoscopy prediction model. DL, difficult laryngoscopy; NDL, no difficult laryngoscopy.

confusion matrix shows the results of counting the number of matches between the original class of the test set target and the test set class predicted by the model (Fig. 3). The precision for DL prediction was 0.18 (ratio of true DL patients among patients predicted as DL, 11/60) and the recall 0.85 (ratio of patients predicted as DL among patients with true DL, 11/13). The importance of the predictors used in the final model is shown in Figure 4.

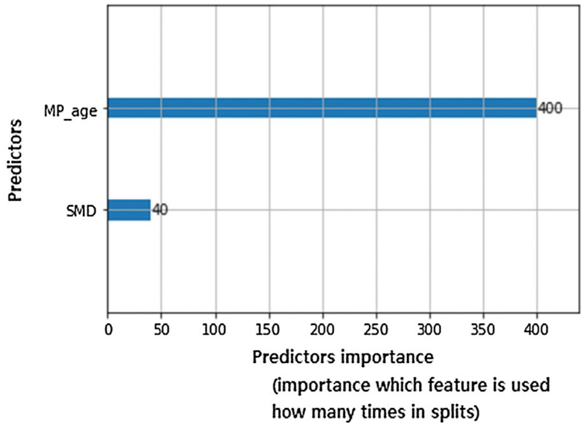

Figure 4 Predictors importance in last model. MP_age, Mallampati grade x age; SMD, sternomental distance.

Discussion

Numerous studies have reported various parameter combinations and results in an attempt to predict DA intubation. In a meta-analysis predicting difficult intubation, the combination of Mallampati score and thyromental distance predicted the most difficult intubation,[17] however, the lack of heterogeneity and the small number of studies limited the conclusions. L'Hermite et al. predicted difficult intubations using the passive scores of five airway parameters but suggested that such predictions of difficult intubation were unlikely to be useful.[18] Some studies have used ML to predict DL or difficult intubation. However, those study results were not validated and evaluated using a test set. Moustafa et al. used ML with a positive predictive value of 76%, a negative predictive value of 76%, and an AUROC of 0.79 for DL.[11] However, even with nine predictive parameters, their cross-validation results were worse than the current study (AUROC, 0.79 vs. 0.99). On their study, Langeson et al. concluded that even though difficult intubation using computer-assisted calculation was predictive, each parameter used for difficult intubation had a poor predictive value.[10] Their ultimate predictive target was difficult intubation, which slightly differs from our predictive target – DL prediction with less variables. They reported significant variables in BMI, age, and Mallampati score, but in our study, the result of cross-validation of models without BMI was better.

Despite the differences in prediction targets, some prediction models of difficult tracheal intubation not employing ML perform better than ours. The validated results for those models range from 0.79 to 0.87 in AUROC.[11] However, previous models used 4 to 9 predictors.[10–14] The Arne model employed seven variables, and showed the highest AUROC among the validated models.[13] Although many predictors can be used for accurate predictions, increasing its number is not practical. In this study, the objective was to create predictive models that perform well with few predictors. Airway prediction with few parameters is a good target for ML because of its power to quickly synthesize and analyze variable combinations. On the other hand, models with high precision and recall are ideal for DL prediction. High precision during DL prediction can save time and money in DA preparation. Recall is even more critical considering the serious problem of unpredictable DL. However, when the test set was applied to our model, it did not show the same result as the cross-validation results, and precision was low. This can be attributed to training data overfitting. In general, the solution to overfitting is collecting more training data.[19] Indeed, the small proportion of patients with DL did not provide a sufficient number of datasets. In such cases of unbalanced data, the data amount becomes even more critical.

In our model, the cross-validation results improved despite reducing the number of predictors. In addition, the model cross-validation results using the newly combined Mallampati grade x age were better than that with Mallampati grade and age as separate predictors. In ML predictions, the model performance can be improved by removing irrelevant information for predictions or by combining features to create potentially new features.[19]

We showed a model that predicts DL with as few predictors as possible, and could identify important predictors for classifying DLs in machine learning models. However, it is difficult to interpret how those predictors behave in machine learning models. Machine learning is often referred to as a black box. Data comes in and decisions come out, but the process between input and output is opaque.[20] In our model, we can see that Mallampati grade x age is the most important predictor in the model, but it is difficult to understand the meaning of the number.

Difficult intubation can often lead to unexpected and complex consequences, harming the patient. In 2015–2018, there was a difference in hospitalization costs in the United States between patients with difficult airway intubation and those who did not, and the median value was about $20,000 higher in patients with difficult airway intubation.[21] Difficult laryngoscopy is different from difficult tracheal intubation but may cause difficult tracheal intubation. In some studies, half of the difficult laryngoscopy reported a high intubation difficult scale $\geq$ 5 points.[22] Appropriate difficult laryngoscopy prediction can help to anticipate difficult airways and be an effective way to reduce medical costs.

Our study has some limitations. First, similar to previous machine learning studies, it does not apply to pediatric patients as it applies to adult populations. Second, because both training and testing were conducted on Asian subjects, the results could differ in other races. For example, Asian populations have statistically different dimensions from Caucasian populations for chin arch, face length and nose protrusion.[23] In order to generalize to global anesthesia patients, studies with additional variables may be needed. Recently, video laryngoscopes have been used in difficult airways, so difficult airway management is better than in the past. However, in situations where video laryngoscopes are not available, difficult laryngoscope prediction can be useful. Third, the use of video laryngoscope or fiberoptic endoscope for tracheal intubation was excluded in our study. It reduces the incidence of difficult laryngoscopy results, which may limit the variability of predictors.

In conclusion, the proposed novel model can predict DL with only three predictors (Mallampati grade, age, and sternomental distance). This small number of predictors makes the prediction simple and with high recall. Thus, the likelihood of serious problems caused by DL prediction failure is low. However, for the improvement of the model's overall performance and generalizability more data from patients with DLs and from different races is required.

Funding

This research was supported by the First Research in Lifetime Program of the National Research Foundation (NRF) funded by the Korean government (MSIT) (NRF-2018R1C1B5085866), South Korea.

Conflicts of interest

The authors declare no conflicts of interest.

References

1. Apfelbaum JL, Hagberg CA, Caplan RA, et al. Practice guidelines for management of the difficult airwayan updated report by the American Society of Anesthesiologists task force on management of the difficult airway. Anesthesiology. 2013;118:251–70.
2. Workeneh SA, Gebregzi AH, Denu ZA. Magnitude and Predisposing Factors of Difficult Airway during Induction of General Anaesthesia. Anesthesiol Res Pract. 2017;2017:5836397.
3. Etezadi F, Ahangari A, Shokri H, et al. Thyromental height: a new clinical test for prediction of difficult laryngoscopy. Anesth Analg. 2013;117:1347–51.
4. Frerk C. Predicting difficult intubation. Anaesthesia. 1991;46:1005–8.
5. Khan ZH, Kashfi A, Ebrahimkhani E. A comparison of the upper lip bite test (a simple new technique) with modified Mallampati classification in predicting difficulty in endotracheal intubation: a prospective blinded study. Anesth Analg. 2003;96:595–9.
6. Mallampati SR, Gatt SP, Gugino LD, et al. A clinical sign to predict difficult tracheal intubation; a prospective study. Can Anaesth Soc J. 1985;32:429–34.
7. Riad W, Vaez MN, Raveendran R, et al. Neck circumference as a predictor of difficult intubation and difficult mask ventilation in morbidly obese patients: A prospective observational study. Eur J Anaesthesiol. 2016;33:244–9.
8. Savva D. Prediction of difficult tracheal intubation. Br J Anaesth. 1994;73:149–53.
9. Türkan S, Ates Y, Cuhruk H, et al. Should we reevaluate the variables for predicting the difficult airway in anesthesiology? Anesth Analg. 2002;94:1340–4.
10. Langeron O, Cuvillon P, Ibanez-Esteve C, et al. Prediction of Difficult Tracheal IntubationTime for a Paradigm Change. Anesthesiology. 2012;117:1223–33.
11. Moustafa MA, El-Metainy S, Mahar K, et al. Defining difficult laryngoscopy findings by using multiple parameters: A machine learning approach. Egyptian Journal of Anaesthesia. 2017;33:153–8.
12. Wilson M, Spiegelhalter D, Robertson J, et al. Predicting difficult intubation. Br J Anaesth. 1988;61:211–6.
13. Arne J, Descoins P, Fusciardi J, et al. Preoperative assessment for difficult intubation in general and ENT surgery: predictive value of a clinical multivariate risk index. Br J Anaesth. 1998;80:140–6.
14. Naguib M, Malabarey T, AlSatli RA, et al. Predictive models for difficult laryngoscopy and intubation. A clinical, radiologic and three-dimensional computer imaging study. Can J Anaesth. 1999;46:748.
15. Chawla NV, Bowyer KW, Hall LO, et al. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res. 2002; 16:321–57.
16. Wilson DL. Asymptotic properties of nearest neighbor rules using edited data. IEEE Trans Syst Man Cybern. 1972:408–21.

17. Shiga T, Zi Wajima, Inoue T, et al. Predicting Difficult Intubation in Apparently Normal PatientsA Meta-analysis of Bedside Screening Test Performance. Anesthesiology. 2005;103:429–37.
18. Yentis SM. Predicting difficult intubation–worthwhile exercise or pointless ritual? Anaesthesia. 2002;57:105–9.
19. Géron A. Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems. 1st ed Sebastopol (CA): O'Reilly Media, Inc; 2017.
20. The Lancet Respiratory Medicine. Opening the black box of machine learning, 6; 2018. p. 801. Available at: https://www.thelancet.com/journals/lanres/article/PIIS2213-2600(18)30425-9/fulltext. [accessed 11 January 2011].
21. Phillips K, Moucharite M, Wong T, et al. Cost burden associated with difficult intubation in the United States. Trends Anaesth Crit Care. 2020;30:e131.
22. Adnet F, Borron SW, Racine SX, et al. The intubation difficulty scale (IDS) proposal and evaluation of a new score characterizing the complexity of endotracheal intubation. Anesthesiology. 1997;87:1290–7.
23. Zhuang Z, Landsittel D, Benson S, et al. Facial anthropometric differences among gender, ethnicity, and age groups. Ann Occup Hyg. 2010;54:391–402.