Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

The scoring system of airway examination to predict the risks of difficult intubation in adult patients undergoing intubated general anesthesia: Machine learning approach

Nalinee Kovitwanawong
Prince of Songkla University
Aradchaporn Phetsirada
Prince of Songkla University
Chanatthee Kitsiripant
Prince of Songkla University
Maliwan Oofuvong
Prince of Songkla University
Chanon Kongkamol
Prince of Songkla University
Kanakorn Horsiritham
Prince of Songkla University
Karuna Jitpakdee
Prince of Songkla University
Sujitra Thongkam
Prince of Songkla University
Nalinee Noonuan
Prince of Songkla University
Sumidtra Prathep
sumidtra.p@psu.ac.th
Prince of Songkla University

Article

Keywords: difficult intubation, airway, machine learning, scoring system, neural network, random forest, XGBoost

Posted Date: May 11th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9353898/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

Although the incidence of difficult intubation is relatively low, it remains a major concern for anesthesiologists. Individual airway examination parameters often have limited predictive value. However, combining multiple assessments can improve prediction accuracy. This study aimed to develop a scoring system using a combination of airway examination factors to predict the risk of difficult intubation.

Methods

A retrospective cohort study was conducted in adult patients undergoing intubated general anesthesia at Songklanagarind Hospital between June 2024 and August 2025. Predictive factors included age, body mass index (BMI), inter-incisor gap, Mallampati classification, thyromental distance, upper lip bite test, and neck mobility. Data were obtained from the hospital information system and preprocessed using Min–Max normalization. The dataset was divided into training (80%) and testing (20%) sets. Four machine learning algorithms—Logistic Regression, Random Forest, XGBoost, and Neural Network—were developed and evaluated based on sensitivity, specificity, F1-score, and area under the receiver operating characteristic curve (AUC).

Results

A total of 5,595 adult patients were enrolled, with an incidence of difficult intubation of 5.3% (298 patients). Among the models tested, the Neural Network algorithm showed the highest predictive performance, with a sensitivity of 0.83, specificity of 0.39, F1-score of 0.13, and AUC of 0.61 at an optimal threshold of 0.379. This model was subsequently integrated into a **web-based application** (https://difficultairway.dida.psu.ac.th/) to provide clinicians with real-time risk estimation.

Conclusion

The proposed Neural Network–based scoring system demonstrated potential in predicting difficult intubation. The publicly accessible web application allows clinicians to perform quick, data-driven airway risk assessments. This tool may enhance preoperative airway assessment and support anesthesiologists in preparing for difficult airway management.

Trial registration:

the Human Research Ethics Committee, Prince of Songkla University (REC.66-372-18-9) on 1st September 2023.

Introduction

Endotracheal intubation is a fundamental procedure in most surgeries requiring general anesthesia and in the management of patients with acute respiratory failure. Although the incidence of difficult intubation at Songklanagarind Hospital is relatively low (approximately 0.005%), it remains a significant concern(1) due to the potential for serious complications such as hypoxia, airway trauma, and cardiac arrest. Accurate prediction of difficult intubation enables anesthesiologists to prepare appropriate airway management strategies, reduce complications, and improve patient safety.

Physical airway examination remains the cornerstone of predicting difficult intubation. Several assessment techniques are currently used, including the Modified Mallampati classification, thyromental distance, interincisor gap, neck movement, and upper lip bite test. At Songklanagarind Hospital, these five methods are routinely applied. However, each individual test has limited sensitivity and specificity. Therefore, combining multiple airway examination parameters into a scoring system may enhance the reliability and accuracy of prediction(2, 3), and provide a simple, practical tool for patient assessment—even when performed by non-anesthesiologists.

Unlike previous studies that primarily relied on single-variable analysis or conventional statistical models(2, 4–9), this study employed multiple machine learning algorithms—including logistic regression, random forest, XGBoost, and neural network—to develop and compare predictive models for difficult intubation. By evaluating and optimizing the performance of these algorithms, the study aimed to establish the most accurate and practical scoring system based on combined airway examination parameters to support anesthesiologists in preoperative airway assessment.

Methodology

Study design

This study was a single-center retroprospective cohort study conducted in adult patients undergoing intubated general anesthesia at Songklanagarind Hospital between June 2024 and August 2025. Information on patient demographics, baseline characteristics, airway examination findings and laryngoscopic view grade was collected from the hospital information system (HIS) of the Digital Innovation and Data Analytics (DIDA) unit of the Faculty of Medicine, Prince of Songkla University and Songklanagarind Hospital.

The collected dataset was then systematically processed and used to develop and construct a scoring system for predicting difficult intubation. This dataset served as the foundation for training and validating multiple machine learning models—Logistic Regression, Random Forest, XGBoost, and Neural Network—to identify the optimal predictive algorithm for integration into the final scoring system.

This retrospective study was conducted in accordance with the principles of the *Declaration of Helsinki*. The study protocol was reviewed and approved by the Institutional Ethics Committee of the Faculty of Medicine, Prince of Songkla University (REC-67-178-8-1) on May 15, 2024. Informed consent and

withdrawal criteria were not applicable due to the retrospective study design. We have no funding for this study. Clinical trial number is not applicable.

Conceptual Design

All collected data were preprocessed and transformed into an appropriate format for machine learning model development. Numerical variables, such as age, were categorized into clinically relevant ranges (20–40, 40–60, 60–80, and ≥ 80 years). The dataset was normalized using the Min–Max scaling technique to constrain all feature values within the range of [0, 1].

Following preprocessing, the data were randomly divided into training (80%) and testing (20%) subsets. Four supervised machine learning algorithms—Logistic Regression, Random Forest, XGBoost, and Neural Network—were developed using the training set and evaluated using the testing set as shown in Fig. 1.

Each model's performance was compared based on key classification metrics, including sensitivity, specificity, F1-score, and the area under the receiver operating characteristic curve (AUC). The model demonstrating the best overall predictive performance was selected as the final algorithm for the difficult intubation scoring system.

The finalized predictive model was then integrated into a web-based application to provide clinicians with a practical, real-time tool for preoperative airway risk assessment. This web application allows users to input standard airway examination parameters and automatically obtain a predicted risk score, supporting timely decision-making and improved airway management planning.

Participants

The study included adult patients aged 20 years and above who received general anesthesia requiring oroendotracheal intubation performed with a Macintosh laryngoscope blade.

Patients who were intubated before surgery, those with existing tracheostomy tubes, and cases in which intubation was performed using non-conventional techniques (e.g., videolaryngoscopy or fiberoptic intubation) were excluded.

In this study, **difficult intubation** was defined as any intubation that required additional technical assistance (such as a fiber-optic device, intubating laryngeal mask, or videolaryngoscope), assistance from a second anesthesiologist with three or more attempts, a total intubation time exceeding 10 minutes, or a laryngoscopic view of grade 3 or higher.

Sample size calculation

The required sample size was determined using an online calculator (https://wnarifin.github.io/ssc/sssnsp.html), based on the statistical methodology proposed by **Buderer NM**(10) for incorporating disease prevalence into sensitivity and specificity calculations .

Data from a previous retrospective study by Nalinee et al., involving adult patients who underwent general anesthesia with endotracheal intubation at Songklanagarind Hospital between 2014 and 2021, reported an incidence of difficult intubation of 0.005%, with the sensitivity and specificity of the prediction model being 0.62 and 0.69, respectively.

Using these parameters with a 95% confidence interval and an anticipated dropout rate of 10%, the final calculated sample size required for this study was 5,029 patients. This sample size was considered sufficient to achieve reliable estimation of model performance in predicting difficult intubation within the study population.

Statistical analysis

Model performance was evaluated on the independent testing dataset (20%) with four standard classification metrics: Sensitivity, Specificity, F1-Score, and the area under the receiver operating characteristic curve (AUC-ROC).

Sensitivity (or Recall) was used to quantify the proportion of actual positive cases that were correctly identified. Specificity was used to measure the proportion of actual negative cases that were correctly identified. The F1-Score, the harmonic mean of precision and sensitivity, provided a balanced measure of the model's predictive ability.

The AUC-ROC was selected as the primary metric for final model comparison, as it reflects the model’s overall discriminatory performance across various probability thresholds. For each algorithm—Logistic Regression, Random Forest, XGBoost, and Neural Network—the AUC-ROC, sensitivity, specificity, and F1-score were computed on the testing dataset to determine comparative performance.

Model development

All data processing, model development, and analysis were conducted using Python (version 3.11.10). The primary machine learning models were implemented using Scikit-learn (version 1.5.2), XGBoost (version 3.0.0), and TensorFlow (version 2.1.8). To ensure a repeatable process across all experiments, a fixed random seed of 42 was set.

The Logistic Regression and Random Forest models, sourced from the Scikit-learn package, were specifically parameterized to address the issue of class imbalance in the dataset:

Logistic Regression: The class weight parameter was set to 'balanced' to automatically adjust weights inversely proportional to class frequencies. The solver was set to 'liblinear' for efficient classification.

Random Forest: The class weight parameter was also set to 'balanced' to mitigate the impact of imbalance during tree construction.

For the XGBoost model, the scale_pos_weight parameter was employed to handle class imbalance. This parameter was set to scale_factor, which was defined as the ratio of negative to positive instances (neg_count/pos_count). Additionally, the eval_metric was set to 'logloss' to monitor the model's performance on the validation set during training, a standard practice for assessing predictive performance.

Figure 2 shows the Neural Network (NN) model utilized a Multilayer Perceptron (MLP) architecture. It consisted of 12 fully connected (dense) hidden layers, each containing eight nodes. The Rectified Linear Unit (ReLU) activation function was applied to all hidden layers. The final output layer contained a single node and employed the Sigmoid function to produce a probability score for binary classification. To address the imbalanced nature of the dataset, the model was trained using adjusted class weights. The optimization was performed using the Adam Optimizer (Adaptive Moment Estimation), and the objective loss function was Binary Cross-Entropy. The model was trained for a maximum of 200 epochs. A validation set, comprising 30% (0.3 rate) of the training data, was used for internal performance monitoring. To mitigate overfitting, two primary regularization strategies were employed: A Dropout layer with a rate of 0.2 was inserted between layers 7 and 8. Early stopping was implemented to halt training if the validation loss did not decrease for 10 consecutive epochs. This mechanism was activated after the 100th epoch.

Deployment as a Web Application

The best-performing model, identified through comparative evaluation of AUC-ROC, sensitivity, specificity, and F1-score, was selected for deployment. This final model was integrated into a web-based clinical tool accessible at https://difficultairway.dida.psu.ac.th/. The web application enables clinicians to input standard airway assessment parameters and receive an instant predicted risk score for difficult intubation. This integration provides a real-time decision-support system to enhance preoperative airway evaluation and improve patient safety.

Results

Patient characteristics

A total of **5,595 adult patients** were included in the study, with an overall incidence of **difficult intubation of 5.3% (298 patients)**. Demographic data, baseline characteristics, and airway examination findings were analyzed to identify trends associated with difficult airway predictors.

As shown in Table 1, patients with difficult intubation were generally older than those without difficult intubation. A higher proportion of patients aged ≥ 60 years and ≥ 55 years was observed in the difficult

intubation group. Several airway-related parameters, including inter-incisor gap, limited neck mobility, and certain Mallampati and thyromental distance categories, showed statistically significant differences between groups.

Table 2 presents the comparison after clinically relevant regrouping of variables. Age remained significantly associated with difficult intubation, particularly in patients aged ≥ 55 years. Limited neck flexion–extension also demonstrated a higher prevalence in the difficult intubation group. In contrast, several commonly used airway assessment parameters, such as Mallampati classification, thyromental distance, and upper lip bite test, did not show strong individual associations when evaluated independently.

These findings suggest that while some demographic and airway variables differ between groups, no single parameter demonstrated sufficient predictive strength on its own, supporting the rationale for combining multiple airway assessments within a machine learning–based predictive model.

Table 1
Comparison of All Parameters Among Status

	Case/Control			
	Overall (N = 5595)	**No (N = 5297)**	**Yes (N = 298)**	**P-value**
Age (years)				< 0.001***[1]
Mean (SD)	53.62 (16.37)	53.44 (16.48)	56.94 (13.96)	
Min, Max	20.00, 99.00	20.00, 99.00	20.00, 88.00	
Group of age (n%)				< 0.001***[3]
20–40	1272 (22.73%)	1233 (23.28%)	39 (13.09%)	
40–60	2064 (36.89%)	1954 (36.89%)	110 (36.91%)	
60–80	1988 (35.53%)	1849 (34.91%)	139 (46.64%)	
≥ 80	271 (4.84%)	261 (4.93%)	10 (3.36%)	
Group of age (n%)				< 0.001***[3]
< 40	1272 (22.73%)	1233 (23.28%)	39 (13.09%)	
40–60	2064 (36.89%)	1954 (36.89%)	110 (36.91%)	
≥ 60	2259 (40.38%)	2110 (39.83%)	149 (50.00%)	
Group of age > = 55 year (n%)	2802 (50.08%)	2620 (49.46%)	182 (61.07%)	< 0.001***[3]
Weight (kg)				< 0.001***[1]
Mean (SD)	63.31 (15.25)	63.18 (15.29)	65.49 (14.36)	
Min, Max	0.80, 172.00	0.80, 172.00	33.50, 133.00	
Height (cm)				< 0.001***[1]

1:Wilcoxon test, 2:t-test, 3:Chi-squared test, 4:Fisher's exact test

	Case/Control			
	Overall	**No**	**Yes**	**P-value**
	(N = 5595)	**(N = 5297)**	**(N = 298)**	
Mean (SD)	160.21 (9.07)	160.10 (8.88)	162.10 (11.92)	
Min, Max	15.90, 190.00	20.20, 190.00	15.90, 183.00	
BMI (kg/m^2)				0.116[1]
Mean (SD)	24.63 (5.95)	24.59 (5.26)	25.49 (13.15)	
Min, Max	3.90, 236.00	3.90, 63.10	13.50, 236.00	
Group of BMI (n%)				0.046*[3]
Underweight	430 (7.69%)	401 (7.57%)	29 (9.73%)	
Normal	1839 (32.87%)	1758 (33.19%)	81 (27.18%)	
Overweight	1042 (18.62%)	995 (18.78%)	47 (15.77%)	
obese1	1610 (28.78%)	1513 (28.56%)	97 (32.55%)	
obese2	674 (12.05%)	630 (11.89%)	44 (14.77%)	
Group of BMI (n%)				0.063[3]
Underweight	430 (7.69%)	401 (7.57%)	29 (9.73%)	
Normal	1839 (32.87%)	1758 (33.19%)	81 (27.18%)	
Overweight-obese	3326 (59.45%)	3138 (59.24%)	188 (63.09%)	
Emergency	438 (7.83%)	420 (7.93%)	18 (6.04%)	0.285[3]
The classification of Mallampati (n%)				0.124[4]
1	1891 (33.80%)	1794 (33.87%)	97 (32.55%)	
2	3474 (62.09%)	3292 (62.15%)	182 (61.07%)	
3	199 (3.56%)	184 (3.47%)	15 (5.03%)	
1:Wilcoxon test, 2:t-test, 3:Chi-squared test, 4:Fisher's exact test				

	Case/Control			
	Overall (N = 5595)	**No (N = 5297)**	**Yes (N = 298)**	**P-value**
4	31 (0.55%)	27 (0.51%)	4 (1.34%)	
The Thyromental distance (n%)				0.097_{4}
1	11 (0.20%)	11 (0.21%)	0 (0.00%)	
2	80 (1.43%)	71 (1.34%)	9 (3.02%)	
3	4892 (87.44%)	4640 (87.60%)	252 (84.56%)	
4	612 (10.94%)	575 (10.86%)	37 (12.42%)	
Interincisor gap (n%)				$0.045^{*}{}_{4}$
1	9 (0.16%)	7 (0.13%)	2 (0.67%)	
2	98 (1.75%)	89 (1.68%)	9 (3.02%)	
3	5342 (95.48%)	5064 (95.60%)	278 (93.29%)	
4	146 (2.61%)	137 (2.59%)	9 (3.02%)	
Upper lip bite test				0.566_{4}
1	4703 (84.06%)	4452 (84.05%)	251 (84.23%)	
2	843 (15.07%)	800 (15.10%)	43 (14.43%)	
3	49 (0.88%)	45 (0.85%)	4 (1.34%)	
Limited Neck flexion extension	118 (2.11%)	100 (1.89%)	18 (6.04%)	$< 0.001^{***}{}_{3}$
1:Wilcoxon test, 2:t-test, 3:Chi-squared test, 4:Fisher's exact test				

Table 2
Comparison of Adjusted Parameters Among Status

Age (years)	**Overall (N = 5595)**	**No (N = 5297)**	**Yes (N = 298)**	**P-value**
				< 0.001* [1]**
Mean (SD)	53.62 (16.37)	53.44 (16.48)	56.94 (13.96)	
Min, Max	20.00, 99.00	20.00, 99.00	20.00, 88.00	
Group of age (n%)				< 0.001*** [3]
< 40	1272 (22.73%)	1233 (23.28%)	39 (13.09%)	
40–60	2064 (36.89%)	1954 (36.89%)	110 (36.91%)	
≥ 60	2259 (40.38%)	2110 (39.83%)	149 (50.00%)	
Group of age > = 55 yrs.old (n%)	2802 (50.08%)	2620 (49.46%)	182 (61.07%)	< 0.001*** [3]
Group of BMI (n%)				0.063 [3]
Normal	1839 (32.87%)	1758 (33.19%)	81 (27.18%)	
Underweight	430 (7.69%)	401 (7.57%)	29 (9.73%)	
Overweight-obese	3326 (59.45%)	3138 (59.24%)	188 (63.09%)	
Emergency	438 (7.83%)	420 (7.93%)	18 (6.04%)	0.285 [3]
The classification of Mallampati (n%)				0.115 [4]
1	1891 (33.80%)	1794 (33.87%)	97 (32.55%)	
2	3474 (62.09%)	3292 (62.15%)	182 (61.07%)	
3	199 (3.56%)	184 (3.47%)	15 (5.03%)	
4	31 (0.55%)	27 (0.51%)	4 (1.34%)	
The Thyromental distance (n%)				0.099 [4]

1:Wilcoxon test, 2:t-test, 3:Chi-squared test, 4:Fisher's exact test

Age (years)	**Overall (N = 5595)**	**No (N = 5297)**	**Yes (N = 298)**	**P-value**
				< 0.001* [1]**
1	11 (0.20%)	11 (0.21%)	0 (0.00%)	
2	80 (1.43%)	71 (1.34%)	9 (3.02%)	
3	4892 (87.44%)	4640 (87.60%)	252 (84.56%)	
4	612 (10.94%)	575 (10.86%)	37 (12.42%)	
Interincisor gap (n%)				0.043* [4]
1	9 (0.16%)	7 (0.13%)	2 (0.67%)	
2	98 (1.75%)	89 (1.68%)	9 (3.02%)	
3	5342 (95.48%)	5064 (95.60%)	278 (93.29%)	
4	146 (2.61%)	137 (2.59%)	9 (3.02%)	
Upper lip bite test				0.549 [4]
1	4703 (84.06%)	4452 (84.05%)	251 (84.23%)	
2	843 (15.07%)	800 (15.10%)	43 (14.43%)	
3	49 (0.88%)	45 (0.85%)	4 (1.34%)	
Limited Neck flexion extension	118 (2.11%)	100 (1.89%)	18 (6.04%)	< 0.001*** [3]
1:Wilcoxon test, 2:t-test, 3:Chi-squared test, 4:Fisher's exact test				

Model Performance

The performance metrics for all four machine learning models—**Logistic Regression, Random Forest, XGBoost, and Neural Network**—are summarized in Table 3. The Neural Network model demonstrated the highest sensitivity (0.83) and AUC (0.61), indicating superior discriminative power in identifying difficult intubations compared with the other algorithms as shown in Fig. 3.

Table 3
Performance metrics for all four machine learning models

Metric	Logistic Regression	Random Forest	XGBoost	Neural Network
Sensitivity	0.491525	0.423729	0.42373	0.830508
Specificity	0.766038	0.772642	0.75755	0.393396
F1-Score	0.172619	0.153846	0.14663	0.130493
AUC	0.59661	0.57504	0.57325	0.614023

The Neural Network model's superior sensitivity reflects its ability to identify high-risk cases, which is desirable in clinical airway management. While specificity was lower, the balance achieved between recall and discriminative ability makes the model well suited as a screening tool for preoperative difficult airway prediction.

Discussion

The descriptive analyses presented in Tables 1 and 2 highlight the limitations of relying on individual airway assessment parameters to predict difficult intubation. Although several variables such as advanced age, limited neck mobility, and inter-incisor gap were more frequently observed in patients with difficult intubation, most parameters demonstrated only modest associations when considered independently.

Notably, commonly used bedside airway assessments, including Mallampati classification and thyromental distance, did not consistently differentiate between difficult and non-difficult intubation groups. These findings are consistent with previous literature(3) demonstrating that individual airway tests have limited sensitivity and specificity when used in isolation.

The observed overlap of airway characteristics between groups underscores the multifactorial nature of difficult intubation and supports the use of multivariable approaches. By integrating multiple airway parameters simultaneously, machine learning models, particularly neural networks, can capture complex, nonlinear interactions that are not apparent in univariate or traditional regression analyses.

Therefore, the results presented in Tables 1 and 2 provide important justification for the application of a combined machine learning–based scoring system rather than reliance on single airway predictors in routine clinical practice.

This study developed and compared four supervised machine learning algorithms—Logistic Regression (LR), Random Forest (RF), XGBoost, and Neural Network (NN)—to predict difficult intubation using routinely preoperative airway parameters. Among these models, the NN algorithm demonstrated the highest sensitivity (0.83) and the largest AUC (0.61), indicating improved ability to identify patients at risk of difficult intubation in a screening context.

The higher sensitivity observed with the NN model may be attributed to its capacity to capture nonlinear and complex interactions among multiple airway parameters including Mallampati classification, thyromental distance, inter-incisor gap, upper lip bite test and neck mobility. Through their multi-layered structure, the NN can recognize subtle relationships between predictors that simpler models like Logistic Regression or Random Forest may fail to detect.

However, the NN model also showed lower specificity (0.39), suggesting a higher rate of false-positive predictions. This trade-off reflects the model's tendency to prioritize sensitivity—detecting most difficult intubation cases—at the expense of over-predicting risk in patients with normal airways. In clinical practice, such a bias may still be acceptable, as failing to anticipate a difficult airway poses greater risk than overestimating it.

In contrast, the Random Forest and XGBoost models exhibited more balanced specificity (≈ 0.76) but lower sensitivity (≈ 0.42), while Logistic Regression maintained the most consistent overall performance with moderate sensitivity and specificity. These findings highlight the potential of Neural Networks to enhance early detection of difficult intubations, though further optimization is required to improve specificity and overall predictive precision.

Several machine learning–based models have been proposed to support airway risk prediction across different clinical settings. However, substantial heterogeneity exists in terms of patient populations, outcome definitions, predictor selection, and intended clinical use, limiting direct comparison of performance metrics across studies.

The Alex Difficult Laryngoscopy Software (ADLS) developed by Moustafa et al.(11) represents one of the earliest applications of machine learning to airway prediction. ADLS employed a decision tree (J48) algorithm trained on 100 matched cases and demonstrated balanced sensitivity and specificity of approximately 76%. In contrast, our NN model was developed using a substantially larger prospective cohort of 5,595 patients, enabling training and evaluation under more heterogeneous, real-world perioperative conditions. Importantly, while ADLS emphasized balanced accuracy, our model was deliberately optimized to prioritize sensitivity in order to function as a screening tool, reflecting the clinical priority of minimizing unanticipated difficult airway events.

Recent studies have demonstrated the potential of neural network–based models for predicting difficult airway management; however, their performance and clinical applicability vary considerably depending on patient population, clinical setting, and optimization objectives.

Zhou et al.(12) applied multiple machine learning algorithms, including neural network–based models, to predict difficult airway intubation in patients undergoing thyroid surgery. Their study reported high discriminative performance, wi th AUC values exceeding 0.8, reflecting the relatively homogeneous and high-risk nature of the study population. Thyroid surgery patients frequently present with airway distortion related to goiter or tracheal deviation, which may enhance model discrimination. However, despite the high AUC, the reported recall (sensitivity) of several models was comparatively lower,

suggesting that model optimization primarily emphasized overall discrimination rather than maximizing case detection. In contrast, our Neural Network model was specifically optimized to prioritize sensitivity, consistent with its intended role as a screening tool in routine perioperative airway assessment across diverse surgical populations.

Yamanaka et al.(13) evaluated machine learning models, including a multilayer perceptron (MLP), in a large multicenter cohort of emergency department intubations. While MLP represents a form of neural network, the clinical context and model objectives differed substantially from those of the present study. Emergency intubations are characterized by physiological instability, time constraints, and marked variability in operator experience, and the primary outcome measures focused on discrimination and first-pass success. Although their models achieved moderate discriminative performance, sensitivity was not the primary optimization target. In contrast, our Neural Network model was developed for the perioperative setting, where preoperative airway assessment allows for proactive risk stratification, and sensitivity is prioritized to minimize the risk of unanticipated difficult airway events.

Taken together, these comparisons highlight that neural network–based models should not be evaluated solely based on summary performance metrics such as AUC. Instead, model performance must be interpreted within the context of clinical intent and application. Whereas prior neural network models emphasized discrimination in specific high-risk or emergency settings, our approach prioritized sensitivity to support early identification and preparedness in routine perioperative practice. This distinction explains the observed differences in performance metrics and underscores the importance of aligning model design with clinical decision-making priorities in airway management.

Observed Incidence of Difficult Intubation

The incidence of difficult intubation observed in this study (5.3%) was higher than previously reported at Songklanagarind Hospital. This finding is likely related to differences in case definition and inclusion criteria. In clinical practice, patients anticipated to have difficult airways are often managed with videolaryngoscopy or fiberoptic intubation and therefore excluded from standard laryngoscopic attempts.

In this study, difficult intubation was defined to include cases with a Cormack–Lehane laryngoscopic view grade of 3 or 4, even when intubation was ultimately successful. This broader definition was intentionally applied to capture patients with potentially challenging airways who may benefit from early recognition and preparedness.

When compared with previous reports, the observed incidence in our cohort is consistent with published data using similar laryngoscopic-based definitions. El-Ganzouri et al.(5) reported an incidence of 5.1% for Cormack–Lehane grade 3 and 1.0% for grade 4 views, which is comparable to the incidence observed in our study. Similarly, Ambesh et al.(8) reported higher incidences of Cormack–Lehane grade 3 (9.0%) and grade 4 (1.6%), reflecting differences in patient populations and airway characteristics.

Zhou et al.(12) reported a higher incidence of difficult intubation (9.6%) using a definition that also incorporated difficult laryngoscopy criteria. This higher incidence may be explained by differences in surgical population, as their study focused on patients undergoing thyroid surgery, who frequently present with airway distortion related to goiter or tracheal deviation. These findings suggest that even when similar definitions are applied, underlying anatomical and population-specific factors substantially influence the reported incidence of difficult airway events.

By incorporating cases with poor laryngoscopic views, the proposed scoring system emphasizes sensitivity and is intended to function as a preventive screening tool in routine perioperative airway management, rather than to estimate the true incidence of failed intubation.

Clinical Application and Future Directions

To enhance clinical utility, the final predictive model was integrated into a web-based application available at https://difficultairway.dida.psu.ac.th/, allowing clinicians to input preoperative airway parameters and instantly obtain an estimated risk score for difficult intubation. The tool aims to support anesthesiologists and other healthcare providers in preoperative airway evaluation and facilitate early identification of patients who may benefit from advanced airway planning. The web-based application interface and access QR code are provided in Additional file 1.

Future research should focus on the external validation across multiple institutions and patient populations. Evaluating its performance in various clinical settings will help ensure generalizability, reproducibility, and reliability before large-scale clinical adoption. Additional improvements may include real-time integration with hospital information systems, user interface optimization, and adaptive model updates using new data to maintain performance over time.

Limitations

This study was conducted in a single tertiary care center, which may limit the generalizability. Patients anticipated to have difficult airways were frequently managed with advanced airway techniques and excluded from conventional laryngoscopic attempts, potentially leading to underrepresentation of high-risk cases. In addition, reliance on procedural records may introduce reporting bias. The inclusion of operator-related factors, additional clinical variables, and multicenter validation may further enhance model performance and applicability.

Conclusion

The Neural Network model demonstrated the highest sensitivity and AUC among the evaluated algorithms, supporting its potential role as a screening tool for identifying patients at risk of difficult intubation. Although specificity was limited, prioritizing sensitivity is clinically appropriate in airway management, where failure to anticipate a difficult airway may result in serious complications.

By integrating the model into a publicly accessible web application (https://difficultairway.dida.psu.ac.th/), this study provides a practical, real-time decision-support tool to assist anesthesiologists in preoperative airway assessment and preparation.

Future research should emphasize external validation of the web application in diverse hospital settings to confirm its accuracy and clinical reliability across different populations. Broader validation and integration into perioperative workflows will be essential steps toward establishing this model as an evidence-based adjunct to routine difficult airway management.

Declarations

There are no competing interests for any author.

Funding declaration: there was no funding

- Ethics approval and consent to participate

This retrospective study was conducted in accordance with the principles of the *Declaration of Helsinki*. The study protocol was reviewed and approved by the Human Research Ethics Committee, Prince of Songkla University (REC-67-178-8-1). Due to the retrospective nature of the study and the use of anonymized data, the requirement for informed consent was waived by the Human Research Ethics Committee, Prince of Songkla University.

● Consent for publication

The consent for publication is not applicable. Patients or the public were not involved in the design, or conduct, or reporting, or dissemination plans of our research.

- Availability of data and materials

The university database was recorded in the information system of the Digital Innovation and Data Analytics (DIDA) unit of the Faculty of Medicine, Prince of Songkla University and Songklanagarind Hospital. The datasets generated and analyzed during the current study are not publicly available due to reasons of participant privacy, institutional policy and data use agreements, but be available from the corresponding author on reasonable request and with appropriate ethical approval. We could contact Kanakorn.h@psu.ac.th; data analytics (happy@dida.psu.ac.th) to obtain access to the raw data analysed in this study.

- Competing interests
    - There are no competing interests of all authors.
- Funding
    - There was no funding for this study.

- Authors' contributions

N.K., A.P., C.K. and S.P. draft the idea of proposal

A.P., K.J., S.T., N.N. and C.K. data collection and cleaning the data

C.K and K.H. analyze the data

N.K., A.P., C.K., M.O., C.K., K.H. and S.P. manuscript writing

● Acknowledgements

Not applicable

References

1. Plansangkate P, Wasinwong W, Charuluxananan S, Lapisatepun W, Sriraj W, Pitimana-aree S, et al. Anticipated and unanticipated difficult intubation in the perioperative and anesthetic adverse events in Thailand (PAAd Thai) study. J Med Assoc Thai. 2019;102(2):156-63.
2. Eberhart LHJ, Arndt C, Aust HJ, Kranke P, Zoremba M, Morin A. A simplified risk score to predict difficult intubation: development and prospective evaluation in 3763 patients. Eur J Anaesthesiol. 2010;27(11):935-40.
3. Detsky ME, Jivraj N, Adhikari NK, Friedrich JO, Pinto R, Simel DL, et al. Will this patient be difficult to intubate? The rational clinical examination systematic review. JAMA. 2019;321(5):493-503.
4. Wilson ME, Spiegelhalter D, Robertson JA, Lesser P. Predicting difficult intubation. Br J Anaesth. 1988;61:211-6.
5. El-Ganzouri AR, McCarthy RJ, Tuman KJ, Tanck EN, Ivankovich AD. Preoperative Airway Assessment: Predictive Value of a Multivariate Risk Index. Anesth Analg. 1996;82:1197-204.
6. Naguib M, Scamman FL, O'Sullivan C, Aker J, Ross AF, Kosmach S, et al. Predictive performance of three multivariate difficult tracheal intubation models: a double-blind, case-controlled study. Anesth Analg. 2006;102:818-24.
7. L'Hermite J, Nouvellon E, Cuvillon P, Fabbro-Peray P, Langeron O, Riparta J. The simplified predictive intubation difficulty score: a new weighted score for difficult airway assessment. Eur J Anaesthesiol. 2009;26(12):1003-9.
8. Ambesh SP, Singh N, Rao PB, Gupta D, Singh PK, Singh U. A combination of the modified Mallampati score, thyromental distance, anatomical abnormality, and cervical mobility (M-TAC) predicts difficult laryngoscopy better than Mallampati classification. Acta Anaesthesiol Taiwan. 2013;51:58-62.
9. Arné J, Descoins P, Fusciardi J, Ingrand P, Ferrier B, Boudigues D, 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.

10. Buderer NMF. Statistical methodology: I. Incorporating the prevalence of disease into the sample size calculation for sensitivity and specificity. Acad Emerg Med. 1996;3:895-900.
11. Moustafa MA, El-Metainy S, Mahar K, Abdel-magied EM. Defining difficult laryngoscopy findings by using multiple parameters: A machine learning approach. Egypt J Anaesth. 2017;33:153-8.
12. Zhou C-M, Wang Y, Xue Q, Yang J-J, Zhu Y. Predicting difficult airway intubation in thyroid surgery using multiple machine learning and deep learning algorithms. Front Public Health. 2022;10:937471.
13. Yamanaka S, Goto T, Morikawa K, Watase H, Okamoto H, Hagiwara Y, et al. Machine Learning Approaches for Predicting Difficult Airway and First-Pass Success in the Emergency Department: Multicenter Prospective Observational Study. Interact J Med Res. 2022;11(1):e28366.

Figures

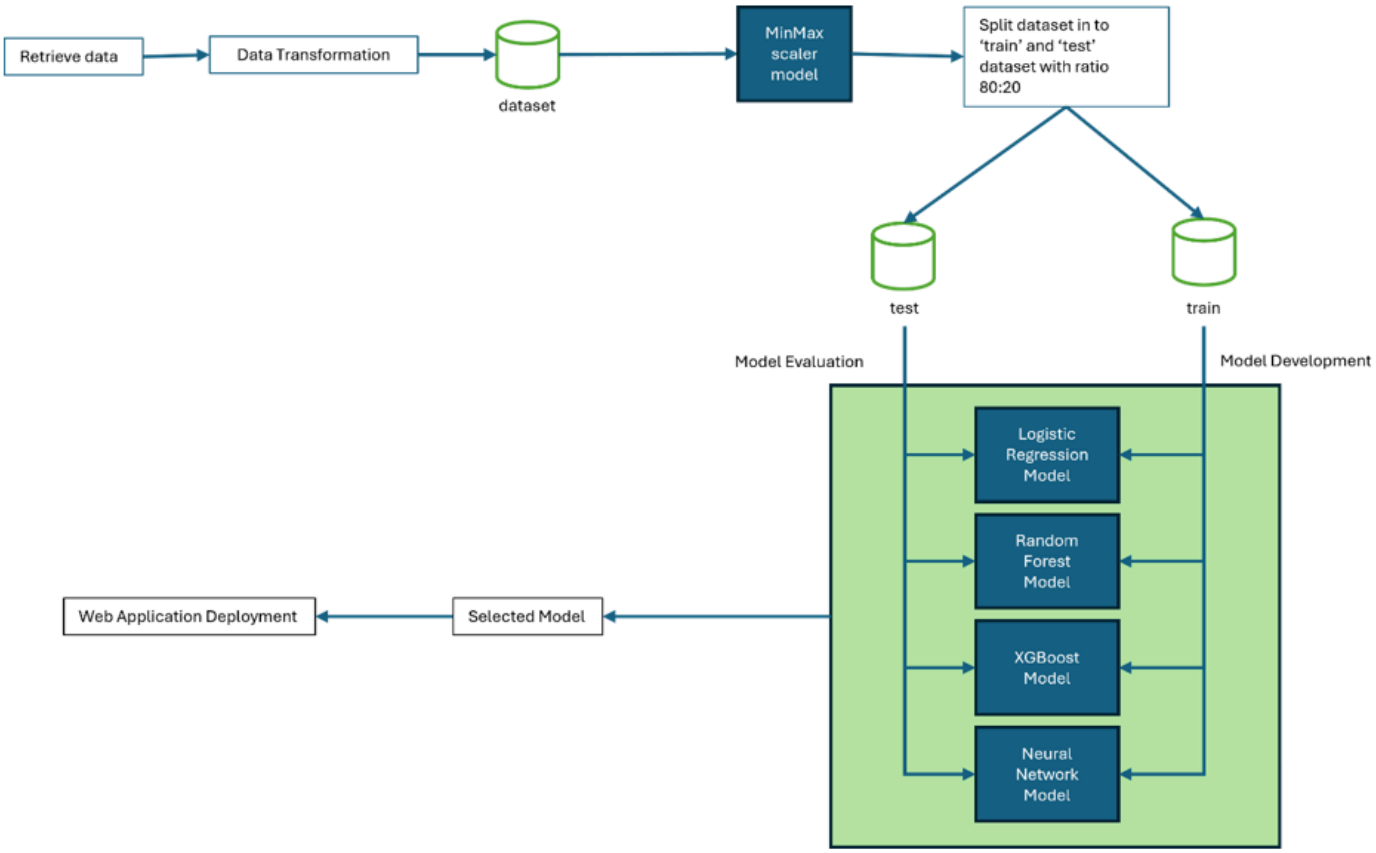


Figure 1

Conceptual Design

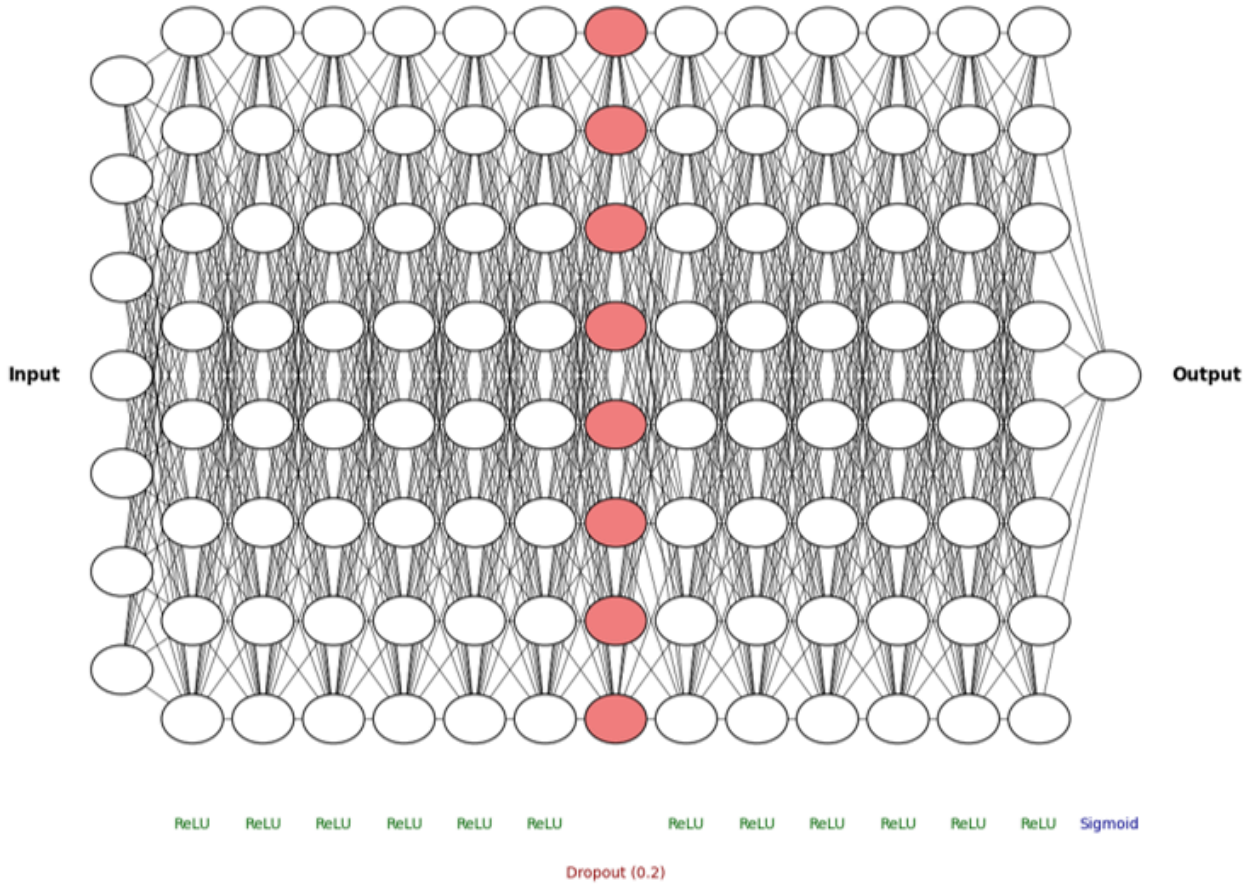


Figure 2

The Neural Network Algorithm

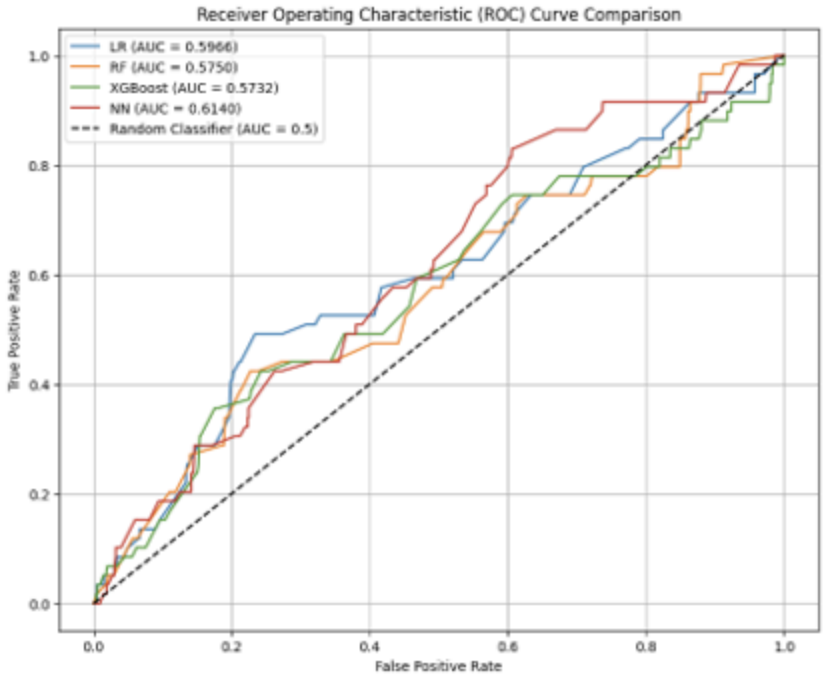


Figure 3

The AUC-ROC Comparison

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1.docx