Archives of Anesthesiology and Critical Care (In Press); x(x): xx-xx.

Available online at http://aacc.tums.ac.ir

Research Article

The Future of Airway Management: A Comparative Analysis of Machine Learning and Ensemble Algorithms for Predicting Associated Factors of Difficult Intubation

Senobar Naderian[1], Mojtaba Mohammadzadeh Lame[2], Ata Mahmoodpoor[2], Arezoo Nejabatian[3], Kavous Shahsavarinia[4], Maryam Soleimanpour[5], Hassan Soleimanpour[6]*

[1]Student Research Committee, Tabriz University of Medical Sciences, Tabriz, Iran.

[2]Department of Anesthesiology and Critical Care Medicine, Tabriz University of Medical Sciences, Tabriz, Iran.

[3]Department of Emergency Medicine, Tabriz Medical Science, Islamic Azad University, Tabriz, Iran.

[4]Research Center for Evidence-Based Medicine, Iranian EBM Centre: A JBI Centre of Excellence, Tabriz University of Medical Sciences, Tabriz, Iran.

[5]Clinical Research Development Unit of Tabriz Valiasr Hospital, Tabriz University of Medical Sciences, Tabriz, Iran.

[6]Medical Philosophy and History Research Center, Imam Reza General Hospital, Tabriz University of Medical Sciences, Tabriz, Iran.

ARTICLE INFO

Article history:
Received 09 March 2025
Revised 30 March 2025
Accepted 14 April 2025

Keywords:
Difficult intubation;
Machine learning;
Ensemble learning;
Mallampati score;
Upper lip bite test;
Feature-selection

ABSTRACT

Background: This study aimed to evaluate the effectiveness of machine learning (ML) models in predicting difficult intubation among maxillofacial surgery patients by using clinical data from a previous study involving 132 patients. The study sought to enhance anesthesiologists' ability to identify patients at risk of difficult intubation, a critical concern in surgical settings.

Methods: The research applied various ML algorithms, including decision trees (DT), random forests (RF), Naive Bayes (NB), neural networks (NN), support vector machines (SVM), K-nearest neighbors (KNN), and ensemble voting methods, to the existing clinical dataset. This dataset contained a range of factors potentially associated with DI, such as the Mallampati score, Upper Lip Bite Test (ULBT) results, facial angle, and other relevant variables. A comprehensive approach was taken to explore the impact of different data preprocessing techniques, with a particular focus on feature selection and normalization methods.

Results: The study found that the combination of mutual information-based feature selection and robust scaler normalization consistently yielded high predictive accuracy. Notably, the decision tree algorithm achieved an accuracy of 0.84 and precision, sensitivity, and specificity scores of 0.95. The analysis also highlighted the strength of ensemble learning, which, by combining multiple classifiers, achieved an accuracy of 0.82. The results suggest that ML models, especially random forests and ensemble voting methods, can be highly accurate in predicting difficult intubation when trained on existing clinical data.

Conclusion: The research underscores the importance of data preprocessing in enhancing algorithmic performance, particularly the effectiveness of mutual information-based feature selection combined with robust scaler normalization. However, the study also indicates the need for further research to refine these models, ensuring their applicability and reliability in real-world clinical settings.

The authors declare no conflicts of interest.
*Corresponding author.
E-mail address: h.soleimanpour@gmail.com

Introduction

Anesthesiologists recognize the critical importance of managing difficult airways in clinical practice [1]. According to the American Society of Anesthesiologists (ASA) Task Force, difficulties with mask ventilation, laryngoscopy, or tracheal intubation signify challenging airway management, each with its specific criteria for difficulty [2]. The ASA defines difficult tracheal intubation as a situation that takes longer than 10 minutes or more than three tries to place the tracheal tube correctly [3]. Proactive measures, including preoperative airway exams, are essential in identifying patients with problematic airways [4]. Despite their ability to forecast many cases [5], these measurements may not always predict difficulties, underscoring the need for continued evaluation [6]. Difficult tracheal intubations can lead to significant morbidity and mortality cases rate, ranging from 1.5 to 13%. [7]. To address these challenges, several airway devices and techniques, including fiberoptic devices and laryngeal mask airways, have been developed [8]. The Cormack and Lehane classification method, which ranges from Class I (clear view) to Class IV (limited view), helps categorize difficulties encountered during laryngoscopy [9].

The classification system helps classify the challenges encountered during laryngoscopy. Despite these advancements, no universal test has proven effective in predicting all cases of difficult intubation (DI) [9]. Commonly used assessment techniques include the Lemon method, Upper Lip Bite Test (ULBT), Modified Mallampati Score (MMS), neck movement, body mass index (BMI), hyomental distance, thyromental distance (TMD), palm print, head extension, and jaw protrusion [7, 10]. Multiple tests are generally preferred over single assessments due to their limited predictive accuracy [11]. Comprehensive airway evaluation involves both visible and hidden anatomical components, such as the tongue base and larynx [2]. Mallampati classes III and IV, along with reduced facial angle (≤82.5°), have been identified as significant predictors of difficult airway management [8, 12-14]. Other parameters such as TMD, sternomental distance (SMD), interincisor distance (IID), and reduced atlanto-occipital have also been implicated in predicting DI [15].

Artificial intelligence (AI) has increasingly been applied to tasks such as disease detection, screening, and treatment [16-17]. AI excels in managing large datasets and enhancing prediction accuracy through methods like feature extraction [18-20]. Ensemble approaches combine predictions from multiple models to improve overall accuracy and reduce generalization errors [19, 21]. The ensemble behaves and produces results as though it were a single model, even though it consists of several underlying models [20]. Notable methods in this field include bagging, boosting, stacking, and voting [21]. Voting is one of the ensemble algorithms that is utilized the most and increases accuracy and robustness by combining predictions from separate models [22].

Recent studies have explored the application of AI in DI prediction. Wang et al. developed a semi-supervised deep-learning model for difficult airway assessment. The model achieved an accuracy of 90.00% and an Area Under Curve (AUC) of 0.94 [23]. Kim et al. developed a predictive model for difficult laryngoscopy and identified the Mallampati score, age, and sternomental distance as predictive parameters; they achieved a predicted AUC of 0.71 and recall (sensitivity) of 0.85 [24]. Yamanaka et al. developed ML models using demographic and initial airway assessment data for predicting difficult airway and first-pass success in the emergency department. They were able to surpass conventional methods in terms of discrimination ability [25].

Zhou et al. identified age, sex, weight, height, and BMI as the top weighting factors for predicting difficult airways. They achieved an AUC > 0.8, accuracy > 90%, and precision of 100% using the gradient boosting algorithm [26]. Tavolara et al. developed a deep-learning model using facial images to identify difficult-to-intubate patients; they achieved an AUC of 0.7105, leveraging the robust features of multiple face regions for classification [22].

Our study introduces a systematic approach for predicting difficult intubation using a comprehensive dataset from patients undergoing general anesthesia for maxillofacial surgery in Iran. This local dataset is unique as it includes a specific facial angle measurement. By integrating multiple pre-operative assessments, such as the ULBT, Modified Mallampati Test, and cephalometric X-ray, with demographic and clinical data, our study offers an overview of factors influencing DI prediction. Additionally, this is the first time an ensemble model has been applied to this type of data, as previous studies have only utilized ensemble methods on facial images in this era. Our study aims to compare the effectiveness of ensemble learning techniques with varied machine learning (ML) functions, aiming to enhance the accuracy and robustness of the prediction. This approach addresses the research gap in using diverse datasets and aims to improve patient safety and understandings in airway management.

Research questions

- How do various machine learning algorithms, including ensemble models, differ in predicting difficult intubation?
- What are the significant predictors of difficult intubation identified by machine learning models?

Methods

Dataset and Preprocessing

A study by Mahmoodpoor et al. (2012-2013), approved by the Ethics Committee of Tabriz University of Medical Sciences, aimed to identify factors associated with DI in patients undergoing maxillofacial surgeries under general anesthesia [8]. The dataset includes features such as age, facial angle, Cormack-Lehane grades, ULBT, Mallampatic score, sex, BMI, and intubation outcomes, among others. The intubation outcomes were classified as either easy or difficult based on Cormack-Lehane grades. It is necessary to mention that facial angle (FA) in these patients was determined via cephalometry X-ray.

For null and missing data handling purposes, Python packages NumPy and Pandas were used. In this process, we identified missing or empty columns and assigned numerical values to empty cells to represent features, according to the team's expert opinion. The final dataset consisted of 19 columns representing features and targets and 132 rows representing patients. Then, as mentioned previously, the intubation result column, consisting of two classes representing easy and difficult intubation, was selected as the target variable.

Then, three normalization techniques were used for the dataset: robust normalization, min-max, and standard scaler. We decided to use three to be able to assess multiple combinations of these three techniques for feature selection techniques. MinMax_Scaler (MinMax) rescales the data within a specified range, and Standard_Scaler (STD) transforms the data to a mean and standard deviation of 1. Robust_Scaler (Robust) is less accurate in identifying outlier data but rescales the data by eliminating the first quartile. After converting the dataset to a numeric representation, the target was taken out. The target column was then added to each scaled data frame (df) after three scaler data frames were instantiated and suited to the data using the corresponding fit_transform methods. Subsequently, three distinct feature selection methods were used to choose the top five objects for the target variable. The ANOVA F-value (Fscore), chi-square (Chi2), and mutual information (Mutual info) were utilized to assess the significance of each feature. Using each of these three feature selection methods, we were motivated by earlier research on feature selection strategies in medical data analysis. Sikri et al. showed how feature ranking is affected by pre-processing data, which is crucial to meeting the chi-square method's assumptions [27]. Also, a mutual information criterion-based feature selection technique was presented by Sulaiman and Labadin, who demonstrated how well it worked to enhance ML model performance [28]. Furthermore, Hoque et al. presented a greedy feature selection technique based on mutual information theory that showed excellent classification accuracy over a number of datasets [29]. These score functions were used to generate three feature selection objects, which were then fitted to imputed data, and the SelectKBest class from the sklearn.feature_selection module was used to choose the top 5 features. In (Figure 1), a graphic illustration of the comprehensive method is presented.

Afterwards, 6 ML algorithms of Decision Tree (DT), K Nearest Neighbor (KNN), Naïve Bayes (NB), Random Forrest (RF), Support Vector Machine (SVM), and Neural Network (NN) were used to classify the dataset. Developing each of the models consisted of setting a random seed, loading the dataset, extracting the target column, converting features, separating data, determining class weights, training the classifier model, assessing performance with various metrics, and visualizing the outcomes. All were steps involved in each technique. For the target variable, three feature selection techniques were applied, which were followed by three normalization techniques. So finally, nine different combinations of normalization and feature selection techniques were used to assess each algorithm's best performance for the target variables. Accuracy, precision, recall, F1 score, and specificity were among the metrics.

Using the same feature selection and normalization techniques, an ensemble learning algorithm was trained for the target variable in addition to the previously mentioned ML algorithms. Individual base classifiers, such as DT and RF, were trained on the preprocessed dataset using the ensemble learning method, ensuring uniformity in feature selection and normalization across the ensemble. These base classifiers' predictions were then aggregated using a "soft" voting scheme, in which the probabilities predicted by each base classifier are averaged, or a "hard" voting strategy, in which a majority vote decides the final prediction. The performance of the ensemble classifiers was then estimated on unseen data through the use of cross-validation techniques. We computed the accuracy and F1 score and compared them with the results from each classifier separately. Confusion matrices are also produced to offer a thorough examination of the performance of the ensemble model. In Table 1, the algorithm, a combination of each feature selection and normalization method, selected features for each combination, and also performance metrics of each combination are available.

Addressing Research Questions

In line with the overall objective of the study, we used a variety of algorithms and ensemble learning techniques to address RQ 1 regarding the performance of ML algorithms. RQ 2, regarding the identification of important factors associated with DI, was addressed by feature selection techniques, which determined the most significant features contributing to the prediction of DI. Also, in (Figure 2), the schematic combination between normalization and feature selection methods that lead to the final results is presented.

Figure 1- Illustration of the comprehensive method

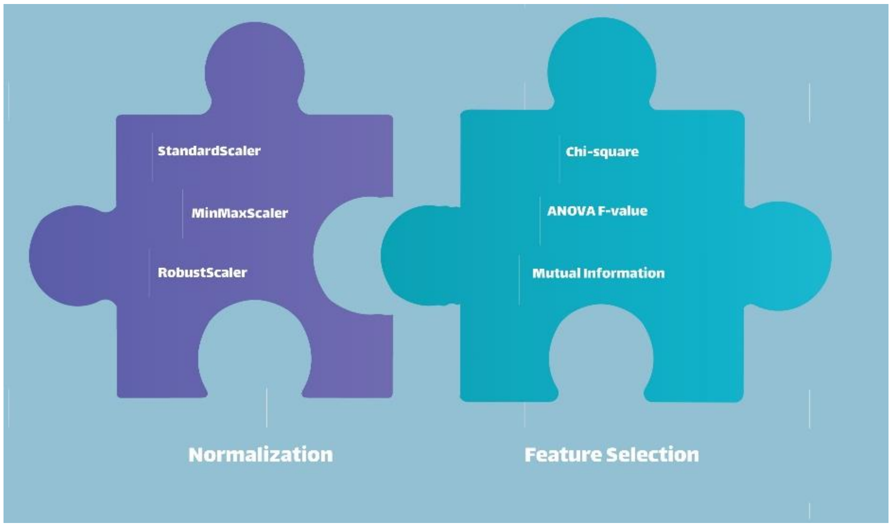

Figure 2- Combination between normalization and feature selection methods

Results

As mentioned previously, (Table 1) presents the performance of various ML algorithms in predicting difficulty in intubation, emphasizing the impact of different combinations of feature selection and normalization techniques on predictive accuracy. Across the algorithms tested, DT, RF, Naïve Bayes, Neural Network, Support Vector Machine, KNN, and an ensemble method by voting were employed. These algorithms were evaluated based on key metrics, including accuracy, precision, sensitivity, and specificity.

Comparing the results, the mutual information feature selection method, coupled with robust scaler normalization, consistently produced high accuracy across several algorithms. For instance, with DT, this combination yielded an accuracy of 0.84, precision, sensitivity, and specificity of 0.95. Similar trends were observed for RF and KNN, achieving accuracies of 0.84 and 0.81 and the other three metrics of 1.0. Additionally, SVM demonstrated competitive performance, particularly with mutual information feature selection and MinMax normalization. With this combination, SVM achieved an accuracy of 0.79, precision of 0.94, sensitivity of 0.95, and specificity of 0.93. This indicates that SVM, when optimized with appropriate feature selection and normalization techniques, can also provide reliable prediction of intubation difficulty.

However, this was not always true among other algorithms. With the mentioned combination, NB exhibited relatively lower performance across different algorithms. For instance, with Chi2 feature selection and MinMax normalization, NB achieved an accuracy of 0.69. Similarly, with Mutual Information feature selection and Robust normalization, NB attained an accuracy of 0.81, precision of 0.83, sensitivity of 0.52, and specificity of 0.89. These results suggest that NB may not be as effective in accurately predicting intubation difficulty as other algorithms tested.

Also, among the algorithms, NN showed its best performance via another combination. At the same time, mutual information was still the choice. The STD scalar combination with it performed extremely well, with an accuracy of 0.81, precision of 0.93, and sensitivity and specificity of 0.82 and 0.96, respectively.

Furthermore, the ensemble method, which included the best and most repeated combinations, namely, mutual information and robust normalization, achieved an accuracy of 0.82, precision of 0.94, sensitivity of 0.78, and specificity of 0.98. This underscores the effectiveness of ensemble learning in combining the strengths of multiple classifiers to enhance predictive performance.

In summary, considering the overall performance, the mutual information feature selection method paired with robust normalization emerges as the most effective combination across multiple algorithms, particularly for DT, RF, and KNN. However, for NB and NN, other combinations, such as Chi2 feature selection with MinMax normalization and mutual information feature selection with standard scalar normalization, respectively, displayed competitive performance. Therefore, the choice of the best combination may depend on the specific algorithm and dataset characteristics.

(Table 2) demonstrates the P values associated with each feature, which indicates their association with binary classes. According to our binary target, each P value is computed based on its association with Class 1, representing easy intubation. The p-value associated with angle input is 0.4, recommending no significant association between this feature and class 1 and also recommending that this feature is more likely to be associated with class 2, DI. 'facial_angle+ULBT' P value is 0.03, showing a significant association between this feature and class 1. The p-value associated with 'facial_angle+mallempati' is 0.02, also suggesting the association with class 1. Intubation_try and BMI P values are 0.3 and 0.6, respectively, indicating that they are more likely to be associated with class 2.

Table 1- Feature Selection, Normalization, and Performance Metrics for Intubation Difficulty Target

Algorithm	Feature Selection	Normalization	Obtained Features from "Feature Selection" and "Normalization" combination	Accuracy	Precision	Sensitivity	Specificity
Decision Tree	Chi2	MinMax	['mallempati', 'facial_angle', 'facial_angle+ULBT', 'facial_angle+mallempati', 'sex']	0.65	0.71	0.65	0.67
		STD	['facial_angle', 'facial_angle+ULBT', 'facial_angle+mallempati', 'sex', 'height']	0.71	0.82	0.82	0.79
		Robust	['facial_angle', 'facial_angle+mallempati+ULBT', 'mallempati', 'sex', 'height']	0.73	0.86	0.86	0.84
	Fscore	MinMax	['facial_angle',	0.71	0.82	0.82	0.79
		STD	'facial_angle+ULBT',	0.71	0.82	0.82	0.79
		Robust	'facial_angle+mallempati', 'sex', 'height']	0.71	0.82	0.82	0.79
	Mutual info	MinMax	['facial_angle',	0.84	0.95	0.95	0.95
		STD	'facial_angle+ULBT',	0.84	0.95	0.95	0.95
		Robust	'facial_angle+mallempati', 'intubation_try', 'bmi']	0.84	0.95	0.95	0.95
Random Forest	Chi2	MinMax	['mallempati', 'facial_angle', 'facial_angle+ULBT', 'facial_angle+mallempati', 'sex']	0.68	0.72	0.60	0.70
		STD	['facial_angle', 'facial_angle+ULBT',	0.75	0.86	0.78	0.87

			'facial_angle+mallempati', 'sex', 'height']				
		Robust	['facial_angle', 'facial_angle+mallempati+ULBT', 'mallempati', 'sex', 'height']	0.74	0.94	0.86	0.96
	Fscore	MinMax	['facial_angle',	0.75	0.86	0.78	0.87
		STD	'facial_angle+ULBT',	0.75	0.86	0.78	0.87
		Robust	'facial_angle+mallempati', 'sex', 'height']	0.75	0.86	0.78	0.87
	Mutual info	MinMax	['facial_angle',	0.84	1.0	1.0	1.0
		STD	'facial_angle+ULBT',	0.84	1.0	1.0	1.0
		Robust	'facial_angle+mallempati', 'intubation_try', 'bmi']	0.84	1.0	1.0	1.0
Naïve Bayes	Chi2	MinMax	['mallempati', 'facial_angle', 'facial_angle+ULBT', 'facial_angle+mallempati', 'sex']	0.69	0.75	0.39	0.82
		STD	['facial_angle', 'facial_angle+ULBT', 'facial_angle+mallempati', 'sex', 'height']	0.78	0.79	0.39	0.88
		Robust	['facial_angle', 'facial_angle+mallempati+ULBT', 'mallempati', 'sex', 'height']	0.74	0.80	0.34	0.90
	Fscore	MinMax	['facial_angle',	0.78	0.79	0.39	0.88
		STD	'facial_angle+ULBT',	0.78	0.79	0.39	0.88
		Robust	'facial_angle+mallempati', 'sex', 'height']	0.78	0.79	0.39	0.88
	Mutual info	MinMax	['facial_angle',	0.81	0.83	0.52	0.89
		STD	'facial_angle+ULBT',	0.81	0.83	0.52	0.89
		Robust	'facial_angle+mallempati', 'intubation_try', 'bmi']	0.81	0.83	0.52	0.89
Neural Network	Chi2	MinMax	['mallempati', 'facial_angle', 'facial_angle+ULBT', 'facial_angle+mallempati', 'sex']	0.78	0.78	0.10	1.0
		STD	['facial_angle', 'facial_angle+ULBT', 'facial_angle+mallempati', 'sex', 'height']	0.76	0.81	0.26	0.96
		Robust	['facial_angle', 'facial_angle+mallempati+ULBT', 'mallempati', 'sex', 'height']	0.79	0.88	0.43	1.0
	Fscore	MinMax	['facial_angle',	0.76	0.74	0.10	1.0
		STD	'facial_angle+ULBT',	0.76	0.81	0.26	0.96
		Robust	'facial_angle+mallempati', 'sex', 'height']	0.78	0.80	0.21	0.97
	Mutual info	MinMax	['facial_angle',	0.80	0.91	0.73	0.95
		STD	'facial_angle+ULBT',	0.81	0.93	0.82	0.96
		Robust	'facial_angle+mallempati', 'intubation_try', 'bmi']	0.80	0.92	0.73	0.97
Support vector Machine	Chi2	MinMax	['mallempati', 'facial_angle', 'facial_angle+ULBT', 'facial_angle+mallempati', 'sex']	0.81	0.79	0.39	0.88

		STD	['facial_angle', 'facial_angle+ULBT', 'facial_angle+mallempati', 'sex', 'height']	0.79	0.79	0.39	0.88
		Robust	['facial_angle', 'facial_angle+mallempati+ULBT', 'mallempati', 'sex', 'height']	0.68	0.74	0.60	0.73
	Fscore	MinMax	['facial_angle', 'facial_angle+ULBT', 'facial_angle+mallempati', 'sex', 'height']	0.82	0.80	0.26	0.95
		STD		0.79	0.79	0.39	0.88
		Robust		0.81	0.80	0.26	0.94
	Mutual info	MinMax	['facial_angle', 'facial_angle+ULBT', 'facial_angle+mallempati', 'intubation_try', 'bmi']	0.79	0.94	0.95	0.93
		STD		0.82	0.81	0.34	0.92
		Robust		0.82	0.83	0.34	0.95
KNN	Chi2	MinMax	['mallempati', 'facial_angle', 'facial_angle+ULBT', 'facial_angle+mallempati', 'sex']	0.76	0.80	0.34	0.90
		STD	['facial_angle', 'facial_angle+ULBT', 'facial_angle+mallempati', 'sex', 'height']	0.79	0.79	0.13	0.99
		Robust	['facial_angle', 'facial angle+mallempati+ULBT', 'mallempati', 'sex', 'height']	0.82	0.75	0.10	0.98
	Fscore	MinMax	['facial_angle', 'facial_angle+ULBT', 'facial_angle+mallempati', 'sex', 'height']	0.81	0.79	0.13	0.99
		STD		0.79	0.79	0.13	0.99
		Robust		0.81	0.79	0.13	0.99
	Mutual info	MinMax	['facial_angle', 'facial_angle+ULBT', 'facial_angle+mallempati', 'intubation_try', 'bmi']	0.74	0.89	0.65	0.95
		STD		0.77	0.89	0.89	0.97
		Robust		0.81	1.0	1.0	1.0
Ensemble of	Mutual info	Robust	['facial_angle', 'facial_angle+ULBT', 'facial angle+mallempati', 'intubation_try', 'bmi']	0.82	0.94	0.78	0.98

Table 2- Feature association with binary classes

Feature	'facial_angle'	'facial_angle+ULBT'	'facial_angle+mallempati'	'intubation_try'	'bmi'
P value	0.4	0.03	0.02	0.3	0.6

Discussion

Our study aimed to develop ML and ensemble learning models based on optimized combinations of feature selection and normalization methods for predicting DI. We used the data of 132 patients who underwent elective maxillofacial surgeries under general anesthesia, including cephalometry, some tests, and demographic data.

Our study's results demonstrate that among various ML algorithms for predicting difficulty in intubation, the mutual information feature selection method, coupled with robust scaler normalization, consistently produced high accuracy across several algorithms. Several studies have shown that the beneficial utility of various feature selection or normalization algorithms, or even a combination of them, is not totally forgotten, even though it may not have been totally reported in our area of interest. As for audiogram analysis, utilizing the various methods or a combination of them resulted in an accuracy of 0.93 in the KNN algorithm [30]. Also, for the aim of enhancing Internet of Things (IOT) botnet attack detection with ML methods, it was shown that using mutual information in combination with other preprocessing necessities, such as normalization, represents accuracy scores that exceed the baseline often [31].

In terms of ML model performance, our study revealed that the RF model emerged as the top performer, exhibiting an accuracy of 0.84 along with flawless precision, sensitivity, and specificity scores of 1.0 each. Balanced RF, which in ‘Balanced’ refers to addressing imbalanced class distribution, had also shown the second highest performance in the prediction of difficult laryngoscopy, with the mean AUC ranging from 0.90 to 0.98 [24] and an accuracy of 0.93 along with a precision of 0.85, sensitivity of 0.73, and specificity of 0.99 in the prediction of preclinical airway management [31].

Following closely behind was the ensemble of five ML models, excluding NN, with an accuracy of 0.82. Despite a slightly lower accuracy compared to RF, this ensemble showcased remarkable precision and specificity. The utilization of ensemble methods has been under investigation recently. In our area of interest, the use of a light gradient boosting machine, which is an ensembled algorithm that sequentially adds the weak gradient boosting to make a stronger prediction of difficult laryngoscopy with an AUC of 0.71 and sensitivity of 0.85 [24]. An ensemble model of ML models also showed a maximum c-statistic of 0.74, with a sensitivity of 0.67 and a specificity of 0.70 among other basic ML models. Predicting difficult airway [25] and an ensemble of convolutional neural networks through majority voting resulted in an AUC of 0.7105 for the prediction of DI [32]. The NN model demonstrated its strength with an accuracy of 0.81, contributing significantly to the ensemble's diversity and resilience. DTs also proved their worth with an accuracy of 0.81 and balanced performance across other metrics, aligning with the ensemble's collective strength. Neural networks have been used in the past few years in the era of airway management, mostly being developed for photographic images with an accuracy of 0.90 [23], for chest X-rays in endotracheal tube (ETT) placement checking with an accuracy of 0.89 [33], and even on numerical datasets for the prediction of DI in thyroid surgery, with an accuracy of 0.90 [26].

KNN, SVM, and NB followed suit. Notably, SVM showcased exceptional specificity, while NB exhibited high precision. However, the SVM model slightly lagged in terms of accuracy compared to other models.

Among the top 5 features related to DI, the Upper Lip Bite Test and Mallampati test score are the most commonly detected and discussed factors, which are the assessment of mouth opening by instructing the patient to bite their upper lip with their lower incisors [34] and the classification grade of visibility of oral and oropharyngeal structures during a maximal mouth opening and tongue protrusion, respectively [35]. As shown in several studies that were undertaken to validate these two factors as predictors for DI, the accuracy of ULBT as a predictor was reported to be up to 0.81%, and for Mallampati, up to 0.66% [36]. Also, these two factors have undergone sensitivity and specificity tests for approval of their prediction roles in DI. The ULBT yielded sensitivity values of 77%, 95.4%, and 75% and specificity scores of 93%, 50.8%, and 54% in three different trials [37-39]. Similarly, two of these studies indicated that the Mallampati test had a sensitivity of 66% and a specificity of 95.5%, with 96% and 54.8%, respectively [37-38]. Also, it should be noted that the combination of these two tests was reported sometimes less consistent than one individual and sometimes more effective than one [36].

Furthermore, in AI-interfered studies, Mallampati was also reported as one of the top five predictors of DI alongside the BMI [24, 26]. The facial angle is one important component that also needs to be investigated. Nonetheless, the absence of research focusing on this particular component suggests the necessity of conducting a thorough investigation in this field. Currently, the study behind the collection of the present study’s dataset serves as the main source of reference [8]. Notably, with a sensitivity of 87.5%, this study demonstrates the importance of the face angle in predicting DI [8]. This discovery highlights the face angle's potential significance as a prognostic marker for DI situations and clarifies its applicability in clinical practice. However, more investigation through other research is necessary to confirm and expand on these results, improving our comprehension of the part that facial anatomy plays in intubation challenges.

Conclusion

To summarize, our research investigated many ML systems that utilize distinct feature selection and normalization techniques to forecast intubation difficulty. With accuracies ranging from 0.81 to 0.84, mutual information feature selection combined with robust scaler normalization proved to be consistently effective, especially when used with DTs, RFs, and K Nearest Neighbors. Certain algorithms performed better than others, although some had very high specificity. Notably, the ensemble robustness and diversity were much enhanced using the neural network model.

Furthermore, we emphasized the significance of particular characteristics, such as the Mallampati score and the Upper Lip Bite Test, in predicting challenging intubations, with studies reporting verified sensitivity and specificity ratings. Furthermore, our analysis highlighted the possible importance of the facial angle, which our dataset's 87.5% sensitivity verified. Limitations include a short dataset size and the requirement for more validation, notwithstanding our findings. However, our work establishes a baseline for further investigation, demonstrating the potential of ML to enhance clinical judgment in airway control.

Abbreviations

ASA: American Society of Anesthesiologists
DI: Difficult Intubation
ULBT: Upper Lip Bite Test
MMS: Modified Mallampati Score
BMI: Body Mass Index
FA: Facial Angle
TMD: Thyromental Distance
SMD: Sternomental Distance
IID: Interincisor Distance
AI: Artificial Intelligence
ML: Machine Learning
AUC: Area Under Curve
MinMax: MinMax_Scaler
STD: Standard_Scaler
Robust: Robust_Scaler
Fscore: ANOVA F-value
Chi2: Chi-square
Mutual info: Mutual Information
DT: Decision Tree
KNN: K Nearest Neighbor
NB: Naïve Bayes
RF: Random Forrest
SVM: Support Vector Machine
NN: Neural Network
IOT: Internet of Things
ETT: EndoTracheal Tube

Ethics approval and consent to participate

Before participation, all individuals provided informed consent via an electronic consent form. This study was approved by the Ethics Committee of Tabriz University of Medical Sciences (date: 2012/6/5, President of Ethics Committee, Dr. Ostadrahimi, Protocol Number: 91107). The study was done anonymously in accordance with the World Medical Association Declaration of Helsinki.

Data availability:

Some sample code used in this paper is available by visiting the GitHub repository that is linked here: https://github.com/senonaderian/difficult_intubation.git.

Also, the dataset will be available via the corresponding author.

Acknowledgment

This article is based on a dataset forming part of Arezoo Nejabatian's specialty thesis. It was registered at Tabriz University of Medical Sciences (No: 92/3-3/2). We would like to appreciate the cooperation of the Clinical Research Development Unit, Imam Reza General Hospital, Tabriz, Iran, in conducting this research.

References

[1] Al Senani Y, Al Shammery AJ, Al Nafea A, Al Absi N, Al Kadhi O, Al-Shammery D. Influence of Fixed Orthodontic Therapy on Pharyngeal Airway Dimensions after Correction of Class-I, -II and -III Skeletal Profiles in Adolescents. Int J Environ Res Public Health, 2021. 18(2):517.

[2] Zhou C, Wang X, Jiang H, Xu H. [The efficacy of using cephalometrics to predict difficult intubation of patients with obstructive sleep apnea hypopnea syndrome]. Shanghai Kou Qiang Yi Xue. 2019;28(3):317-20. Chinese.

[3] Nout E, Bannink N, Koudstaal MJ, Veenland JF, Joosten KF, Poublon RM, et al. Upper airway changes in syndromic craniosynostosis patients following midface or monobloc advancement: correlation between volume changes and respiratory outcome. J Craniomaxillofac Surg. 2012; 40(3):209-14.

[4] Harila-Kaera V, Grön M, Heikkinen T, Alvesalo L. Sagittal occlusal relationships and asymmetry in prematurely born children. Eur J Orthod. 2002;24(6):615-25.

[5] Gupta S, Sharma KR, Jain D. Airway assessment: predictors of difficult airway. Indian J Anaesth. 2004;49(4):257-62.

[6] Costa F, Robiony M, Toro C, Sembronio S, Polini F, Politi M. Simultaneous functional endoscopic sinus surgery and esthetic rhinoplasty in orthognathic patients. J Oral Maxillofac Surg. 2008;66(7):1370-7.

[7] Balafar M, Pouraghaei M, Paknezhad SP, Abad SNA, Soleimanpour H. Evaluation the quality of bag-mask ventilation by E/C, T/E and hook technique (a new proposed technique). BMC Anesthesiol. 2023; 23(1):384.

[8] Mahmoodpoor A, Soleimanpour H, Golzari SE, Nejabatian A, Pourlak T, Amani M, et al. Determination of the diagnostic value of the modified Mallampati score, upper lip bite test and facial angle in predicting difficult intubation: a prospective descriptive study. J Clin Anesth. 2017;37:99-102.

[9] Luckscheiter A, Zink W, Lohs T, Eisenberger J, Thiel M, Viergutz T. Machine learning for the prediction of preclinical airway management in injured patients: a registry-based trial. Clin Exp Emerg Med. 2022; 9(4):304-313.

[10] Mahmoodpoor A, Soleimanpour H, Nia KS, Panahi JR, Afhami M, Golzari SE, et al. Sensitivity of palm print, modified Mallampati score and 3-3-2 rule in prediction of difficult intubation. Int J Prev Med. 2013;4(9):1063-9.

[11] Zhou C, Wang X, Jiang H, Xu H. [The efficacy of using cephalometrics to predict difficult intubation of patients with obstructive sleep apnea hypopnea syndrome]. Shanghai Kou Qiang Yi Xue. 2019; 28(3):317-320.

[12] Urbizu A, Ferré A, Poca MA, Rovira A, Sahuquillo

J, Martin BA, et al. Cephalometric oropharynx and oral cavity analysis in Chiari malformation Type I: a retrospective case-control study. J Neurosurg. 2017; 126(2):626-633.

[13] Kohjitani A, Miyawaki T, Miyawaki S, Nakamura N, Iwase Y, Nishihara K, et al. Features of lateral cephalograms associated with difficult laryngoscopy in Japanese children undergoing oral and maxillofacial surgery. Paediatr Anaesth. 2013;23(11):994-1001.

[14] Soleimanpour M, Rahmani F, Morteza Bagi HR, Ala A, Mahmoodpoor A, Hassani F, et al. Comparison of three techniques on facility of bag-mask ventilation: Thenar eminence, E-O and E-C. Anesth Pain Med. 2018;8(4):e74226.

[15] Kim J, Im KS, Lee JM, Ro J, Yoo KY, Kim JB. Relevance of radiological and clinical measurements in predicting difficult intubation using light wand (Surch-lite™) in adult patients. J Int Med Res. 2016; 44(1):136-46.

[16] Amisha, Malik P, Pathania M, Rathaur VK. Overview of artificial intelligence in medicine. J Family Med Prim Care. 2019; 8(7):2328-2331.

[17] Sarker IH. Machine learning: algorithms, real-world applications and research directions. SN Comput Sci. 2021;2(3):160.

[18] Haneef R, Kab S, Hrzic R, Fuentes S, Fosse-Edorh S, Cosson E, et al. Use of artificial intelligence for public health surveillance: a case study to develop a machine Learning-algorithm to estimate the incidence of diabetes mellitus in France. Arch Public Health. 2021 Sep 22;79(1):168.

[19] Hasan KMZ, Hasan M. Performance evaluation of ensemble based machine learning techniques for prediction of chronic kidney disease. ERCICA. 2018; 1(2019): 415-426. Springer Singapore.

[20] Nahar N, Ara F, Neloy MA, Barua V, Hossain MS, Andersson K. A comparative analysis of the ensemble method for liver disease prediction. 2019.

[21] Gong H, Wang M, Zhang H, Elahe MF, Jin M. An explainable AI approach for the rapid diagnosis of COVID-19 using ensemble learning algorithms. Front Public Health. 2022;10:874455.

[22] Tavolara TE, Gurcan MN, Segal S, Niazi MK. Identification of difficult to intubate patients from frontal face images using an ensemble of deep learning models. Comput Biol Med. 2021;136:104737.

[23] Wang G, Li C, Tang F, Wang Y, Wu S, Zhi H, et al. A fully-automatic semi-supervised deep learning model for difficult airway assessment. Heliyon. 2023;9(5):e15629.

[24] Kim JH, Choi JW, Kwon YS, Kang SS. Predictive model for difficult laryngoscopy using machine learning: retrospective cohort study. Braz J Anesthesiol. 2022;72(5):622-8.

[25] 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. 2021;11.

[26] Zhou CM, Wang Y, Xue Q, Yang JJ, Zhu Y. Predicting difficult airway intubation in thyroid surgery using multiple machine learning and deep learning algorithms. Front Public Health. 2022;10.

[27] Sikri A, Singh NP, Dalal S. Chi-square method of feature selection: impact of pre-processing of data. Int J Intell Syst Appl Eng. 2023;11(3s):241-8.

[28] Suleiman M, Labadin J. Feature selection based on mutual information. 2015:1-6.

[29] Zhou H, Wang X, Zhu R. Feature selection based on mutual information with correlation coefficient. Appl Intell. 2022;52(5):5457-74.

[30] Elkhouly A, Andrew AM, Rahim HA, Abdulaziz N, Malek MFA, Siddique S. Data-driven audiogram classifier using data normalization and multi-stage feature selection. Sci Rep. 2023;13(1):1854.

[31] Al-Sarem M, Saeed F, Alkhammash EH, Alghamdi NS. An aggregated mutual information based feature selection with machine learning methods for enhancing IoT botnet attack detection. Sensors (Basel). 2021;22(1).

[32] Xia M, Jin C, Zheng Y, Wang J, Zhao M, Cao S, et al. Deep learning-based facial analysis for predicting difficult videolaryngoscopy: a feasibility study. Anaesthesia. 2024;79(4):399-409.

[33] Brown MS, Wong KP, Shrestha L, Wahi-Anwar M, Daly M, Foster G, et al. Automated endotracheal tube placement check using semantically embedded deep neural networks. Acad Radiol. 2023;30(3):412-20.

[34] Faramarzi E, Soleimanpour H, Khan ZH, Mahmoodpoor A, Sanaie S. Upper lip bite test for prediction of difficult airway: a systematic review. Pak J Med Sci. 2018;34(4):1019-23.

[35] Stutz EW, Rondeau B. Mallampati Score. In: StatPearls. Treasure Island (FL): StatPearls Publishing; 2024.

[36] Wang LY, Zhang KD, Zhang ZH, Zhang DX, Wang HL, Qi F. Evaluation of the reliability of the upper lip bite test and the modified Mallampati test in predicting difficult intubation under direct laryngoscopy in apparently normal patients: a prospective observational clinical study. BMC Anesthesiol. 2022;22(1):314.

[37] Dawood AS, Talib BZ, Sabri IS. Prediction of difficult intubation by using upper lip bite, thyromental distance and Mallampati score in comparison to Cormack and Lehane classification system. Wiad Lek. 2021;74(9 pt 2):2305-14.

[38] Kar S, Senapati LK, Samanta P, Satapathy GC. Predictive value of modified Mallampati test and upper lip bite test concerning Cormack and Lehane's laryngoscopy grading in the anticipation of difficult intubation: a cross-sectional study at a tertiary care hospital, Bhubaneswar, India. Cureus. 2022;14(9):e28754.

[39] Tang X, Dong Z, Xu J, Cheng P, Wang M, Wang B, et al. Observation of the validity of the upper lip bite test in predicting difficult intubation. Sci Rep. 2023;13(1):22160.