Comparative analysis of popular predictors for difficult laryngoscopy using hybrid intelligent detection methods

Xiaoxiao Liu\textsuperscript{a,1}, Colin Flanagan\textsuperscript{b,1}, Jingchao Fang\textsuperscript{b,1}, Yiming Lei\textsuperscript{c}, Launcelot McGrath\textsuperscript{d}, Jun Wang\textsuperscript{c}, Xiangyang Guo\textsuperscript{e}, Jiangzhen Guo\textsuperscript{f,*}, Harry McGrath\textsuperscript{g,**}, Yongzheng Han\textsuperscript{e,*}

\textsuperscript{a} Electronic and Computer Engineering, University of Limerick, Limerick, Ireland
\textsuperscript{b} Department of Radiology, Peking University Third Hospital, Beijing, China
\textsuperscript{c} Research Centre of Digital Hospital Systems, Peking University, Ministry of Education, Beijing, China
\textsuperscript{d} Northern Hospital Melbourne, Australia
\textsuperscript{e} Department of Anaesthesiology, Peking University Third Hospital, Beijing, China
\textsuperscript{f} School of Engineering Medicine, Beihang University, Beijing, China
\textsuperscript{g} Department of Anesthesiology, University Hospital Limerick, Limerick, Ireland

ARTICLE INFO

Keywords:
Difficult laryngoscopy
Anesthesiology
Laryngoscope exposure
Airway management
Machine learning

ABSTRACT

Difficult laryngoscopy is associated with airway injury, and asphyxia. There are no guidelines or gold standards for detecting difficult laryngoscopy. There are many opinions on which predictors to use to detect difficult laryngoscopy exposure, and no comprehensively unified comparative analysis has been conducted. The efficacy and accuracy of deep learning (DL)-based models and machine learning (ML)-based models for predicting difficult laryngoscopy need to be evaluated and compared, under the circumstance that the flourishing of deep neural networks (DNN) has increasingly left ML less concentrated and uncreative. For the first time, the performance of difficult laryngoscopy prediction for a dataset of 671 patients, under single index and integrated multiple indicators was consistently verified under seven ML-based models and four DL-based approaches. The top dog was a simple traditional machine learning model, Naïve Bayes, outperforming DL-based models, the best test accuracy is 86.6%, the F1 score is 0.908, and the average precision score is 0.837. Three radiological variables of difficult laryngoscopy were all valuable separately and combinedly and the ranking was presented. There is no significant difference in performance among the three radiological indicators individually (83.06% vs. 83.20% vs. 83.33%) and comprehensively (83.74%), suggesting that anesthesiologists can flexibly choose appropriate measurement indicators according to the actual situation to predict difficult laryngoscopy. Adaptive spatial interaction was imposed to the model to boost the performance of difficult laryngoscopy prediction with preoperative cervical spine X-ray.

1. Introduction

A plethora of work has shown that forecasting difficult laryngoscopy is a major preoccupation for critical care specialists, and anaesthetists. Difficult laryngoscopy may cause life-threatening complications [1], including brain damage and death, with a declared impact of 1/110 thousand cases within the United Kingdom. Good exposure and visualization of the laryngoscopy has always been regarded as an essential prerequisite for an effective airway management. Exposure to the larynx is not an issue in the majority of cases; but in patients suffering from cervical spondylopathy, difficult laryngoscopy is frequently experienced [2]. Therefore, it is crucial to make thorough preparations and to provide difficult laryngoscopy in cervical spondylopathy patients to minimize detrimental outcomes.

There is no golden standard for diagnosing difficult laryngoscopy. Conventional indicators of difficult laryngoscopy contain small inter-incisor gap, a high Mallampati score, short thyromental distance and so on. However, none of these predictors can achieve high diagnostic...
accuracy [3]. The vertical distance between the hyoid’s highest point and the mandibulae has been identified as an accurate individual indicator for challenging laryngoscopy in patients suffering from cervical spondylopathy [4]. Another predictor, the corner between the second vertebrae cervicale’s inferior edge and the sixth vertebrae cervicale’s bottom edge in neutral position has been shown to have the potential to be an accurate indicator for cervical spondylopathy patients’ difficult laryngoscopy prediction [5]. The atlanto-occipital gap may reflect reduced range of motion and mild atlanto-occipital fusion, and has proven to have the potential ability to detect difficult laryngoscopy [6]. Our study applies the above three predictors to detect challenging laryngoscopy in cases of cervical spondylopathy, the predictive performance under single factor and multiple indicators will be uniformly verified for the first time.

Automation is key to revolutionizing the use of laryngoscopic imaging in clinical research by augmenting activities ranging from image processing to classification. An unparalleled decade of progress within the domain of DL has demonstrated the capacity to automatically and accurately extract and analyse medical data [7]. DL methods are beginning to be increasingly adopted in medical imaging to improve understanding of difficult laryngoscopy. Compared with labour-intensive human efforts, DL-based methods have significant advantages and potential in imaging processing. These deep-learning methods utilize multiple layers for learning how to depict data at multi-level of abstract for image recognition or classification. However, DL-based methods have limitations, including the arguably limited interpretability of models. Before this, the traditional ML algorithm with strong interpretation and simple principle is the main image processing and classifier, but it is rarely mentioned at present. Only a few studies have focused on comparing the performance of deep learning and traditional machine learning in medical image detection and classification [8]. This work will assess and benchmark the effectiveness and precision of DL-based models and ML-based models for predicting difficult laryngoscopy imaging, given that the recent flourishing of DNN has comparative, and in our view, unwarranted, neglect of ML approaches.

Therefore, the performance of difficult laryngoscopy prediction under single index and comprehensive multiple indicators will be uniformly verified. With the rapid development of DNN, the concentration and creativity of ML are becoming less and less, the effectiveness and accuracy of DL and ML-based prediction models for difficult laryngoscopy need to be evaluated and compared.

2. Background

2.1. Difficult laryngoscopy

Most anesthesiologists experience difficult airway management, and improper management can lead to anesthesia-related death. Numerous physical tests have been conducted to identify individuals with an elevated risk for challenging airways. However, commonly used bedside tests are not suitable to the prediction of challenging laryngoscopy. Considering that the best point of view of laryngoscope is when the axis of mouth, pharynx and larynx are closest to each other, it is difficult for anesthesiologists to expose the entire patient’s glottis [9]. This study comprehensively compared and analyzed multiple predictors of neck mobility, including the vertical gap between the hyoid’s highest point and the mandibulae, the atlanto-occipital gap, and the corner between the second vertebrae cervicale’s inferior edge and the sixth vertebrae cervicale’s bottom edge in neutral position, that may be reflective of difficult laryngoscopy in cervical spondylopathy people.

2.2. Artificial Intelligence

Artificial intelligence (AI) is an advanced technology that has the capacity to properly parse extrinsic database, draw lessons, and adapt flexibly to achieve specific goals and tasks [7]. It covers various levels, from low-grade functions like contours identification, to high-grade functions like comprehension of the full view, for example, features extraction from the images results in quantifying image attributes [10]. Several DL-based approaches, such as CNN [11], transfer learning [12, 13], and some ML-based models, such as Random Forests (RF) [14], Logistic Regression (LR) [15], Support Vector Machines (SVM) [16], etc., will be compared in this study to provide an accurate evaluation of difficult laryngoscopy. Figure 1 shows the difference between DL versus ML. It contains three parts: the comparison of traditional ML flow and DL flow, the training process of ML, and the training process of DL.

A primary sub-category of Artificial Intelligence is ML, where ML sparse database and implement decision-making and forecasts [17]. ML consists of supervised (SL), semi-supervised and un-supervised learning (UL). ML can understand modes in clinical imaging by reviewing voxel strength readings and quantifying image characteristics, entitled “radiomic features” [18], via determining the most appropriate mix and constructing classification/regression models [19]. SL is commonly utilized in imagery for categorization [20] when its outputs are categorized, and for regression [21] missions in which the production variables are ongoing. Parametric ML contains Logistic Regression, Random Forest, Naïve Bayes; Non-parametric ML includes perception, K-Nearest Neighbours [22], Support Vector Machines [23] and XGBoost, etc. Another traditional machine learning algorithm, cosine similarity algorithm [24], was applied to remove similar characteristics of attribute vectors, utilizing Minimum Redundancy Maximum Relevance (mRMR) to validate that the optimum subcategorises of characteristics are representative according to the Pearson correlation coefficient. It favours characteristics with smaller correlational numbers for other traits, and greater correlational numbers for difficult laryngoscope categories. Figure 2 shows the difference between baseline ML and transformational ML. Baseline ML processes structured data, such as contour, joint, bone features, specific feature descriptions of chest and head and neck rays, and converts them to 0/1/2, etc. The transformed ML works directly with unstructured data, namely images.

A powerful subclass of ML is DL, a computational structure which allows to learn high complexity functions based on original database [25]. The primary deep learning technology employed is convolutive neuronal networks [26] (CNN), that maintains the hardcoded shift-invariant, a principal characteristic in imaging. A CNN is made from a pile of layers each conducting a particular action, including convolution, pooling, losses computation. Every mid-layer is given the output from the preceding layer [27]. Discriminant CNNs can boost the performance of image classification, by imposing a Spatial Pyramid Pooling on the feature learning, which can enforce the models to be more discriminative.

3. Materials

3.1. Data description

3.1.1. Study design

From June 2016 to December 2021, people undergoing general anesthesia for selective cervical spinal surgeries were enrolled into the study. The eligibility criteria consisted of the following: (a) Between the ages of 20 and 70. (b) Healthy psychological well-being. (c) Radiography and clinic equipment needs to be comprehensive. The drop-out criteria were outlined below: (a) People suffering from respiratory tumors. (b) Patients suffering from serious cervical injury. (c) Cervical spine is unstable. (d) People suffering from adverse medical condition (ASA class 4 or 5). (e) People who anticipate difficulties in mask ventilation. The radiographic database was procured via examining the patients’ case records and gauging measurements using PACS, Picture Archiving, and Communication System. The research received approval from the Medical Ethics Committee of Peking University Third Hospital, Peking University Health Sciences Center, Beijing, China (IRB00006761-2015021). Enlightened consent was received from every patient. This research has
also received approval from the Chinese Clinical Trial Registry (http://www.chictr.org.cn; identifier: ChiCTR-ROC-16008598).

### 3.1.2. Intubation procedure

Noninvasive BP (Blood Pressure), pulse oximeter, HR (Heart Rate), and electrocardiography were routinely monitored prior to surgery. Anesthetics were injected using propofol (2 mg/kg) and sufentanil (0.3 μg/kg). For patients who were unconscious, rocuronium (0.6 mg/kg) was administered as a neuromuscular blocker. Laryngoscopy classification of sufferers in sniffing position were conducted via a superior anesthetist, and has been evaluated based on the Cormack Lehane score during Macintosh laryngoscopy [28]. People who scored 3 or 4 vision were classified as difficult laryngoscopy, while people who graded 1 or 2 vision were classified as simple laryngoscopy. The Macintosh laryngoscopy was conducted by a superior anesthetist who didn’t take part in pre-operative radiological evaluation. Participants with failed Macintosh laryngoscopy were managed based on the Difficult Airway Society (DAS) 2015 guideline [29].

### 3.1.3. Engaging patients

None of the participants participated in measuring radiological database, or in the development of blueprints to outline as well as conduct this work. None of the participants were asked for suggested interpretations. These findings will be communicated with researchers as well as participants.

### 3.2. Pre-processing

A total of 671 laryngoscopy X-ray images were collected and used in this paper, including 548 easy laryngoscope X-rays, and 123 difficult laryngoscope X-ray images. Pre-processing is vital for medical image classification, especially on small datasets. Most AI techniques for medical images are based on SL, using data sets with imaging database as well as data tags (for example, item categories) [7]. In this paper, histograms and labelled images are applied. The pre-processing work contains data segmentation, labelling, data augmentation [30], and data balancing.
3.2.1. Features extraction and image segmentation

Feature extraction is a crucial pre-processing step for image classification. This paper applied feature extraction methods, such as grey scale conversion, binary transformation, skeleton extraction and central axis transformation, and gradient extraction.

Hospital picture segmentation, which identifies the organ and lesion pixel values based on X-ray imaging database, has the capability to effectively analyse images. It involves providing essential details on the forms and spatial information in such images [27]. This study applied on watershed algorithm for image segmentation. The dimension of all images used is 700 x 700 with bit depth of 8. Figure 3 shows the imaging modalities of patients with cervical spondylosis under different data processing methods. In data labelling, three different labelling methods were shown. Data segmentation contains binarization, the skeleton obtained after the central axis transformation, and the gradient and marker information after watershed map based on gradient. In data augmentation, three modalities of the processing image were listed, with different rescaling, resizing, etc.

Figure 4 shows laryngoscope X-ray feature extraction and image segmentation, it is described as follows:

(1) The first picture consists of raw data.
(2) The second picture is obtained by inverting the original image, getting the opposition, format converting from the original BGR to grayscale and then to RGB display format. The contour line of the bone structure looks clearer than the original image.
(3) The third picture is the original image after binarization. The threshold ranges from 90 to 255, and the pixel value 255 becomes 1 after binarization.
(4) The fourth picture is the skeleton obtained after the central axis transformation, and the scattered line segments around mark the distance line between the points on the central axis and the background pixel points.
(5) The fifth picture is the gradient of the image obtained after noise filtering, using gradients below 10 as starting gradient points.
(6) The sixth picture is obtained by the fifth figure, where gradient and marker information are used to obtain watershed map based on gradient.

3.2.2. Labelling

With the emergence of CNN, supervised DL performed at the leading edge in numerous medical imaging classification missions, using images [31] of various modalities [32]. For dataset training, large amounts of images tagged with the categories for pixels or voxels are required. In practice, collecting heavily tagged biomedical pictures is tough as marking these data requires domain specific awareness and pixel-by-pixel annotation may take time [33].

In this paper, three related radiological indicators were labelled (Figure 5). These three labels are:

Label-1: Vertical gap between hyoid’s highest point to mandibulae [4].
Label-2: Atlanto-occipital distance [6].
Label-3: The corner between the lower margin of the second vertebrae cervicale and the lower margin of the sixth vertebrae cervicale in neutral position [5].

All Three labels were based on indicators previously reported in the literature, as well as indicators that might be meaningful considering the clinical experience of professional anesthesiologists.

3.2.3. Data augmentation

Data Augmentation is an cost-effective way to expand the number and variety of dataset through stochastic modelling [34]. within the imaging field, the current increments encompass include rescaling, resizing, or flipping the image horizontally.
In this study, rotation changes, width changes, height changes, random shear, zoom changes, and horizontal flip were applied to carry out data augmentation. For CNN models, the size of each image was resized to 64 × 64 after data augmentation. For traditional machine learning models, the size of each image was resized to 180 × 180.

3.2.4. Balanced dataset

Unbalanced intensity allocation in laryngoscope database can result in unclear division as well as incompatible outcomes [35]. In the research, histograms equalisation, which plays a key role for image quality enhancement [36], was used to reduce these problems [37]. First, it divides the image into different grid areas, limiting histograms to restricted areas to conduct functions across various areas [35]. Then, it briefly evens out the histograms of grey-scale images, standardizes intensity, and enhances contrast.

Class weights were utilised to weight various data classes when training the dataset. It tells the pattern to “pay more attention” to images of underrepresented classes. Weight values are set based on the ratio of difficult and non-difficult laryngoscopy images.

4. Methods

This work applied DL-based models, such as CNN, DenseNet-121, ResNet-50, VGG-16, ML-based models, like random forest, logistic regression, KNN, and cosine similarity algorithm. This dataset was shuffled and split to training, testing and validation sets in a 7:1:2 proportion.

4.1. Multi-label image classification structure

Figure 6 shows the structure of Multi-Label Image Classification (MLIC) applied in this article. This MLIC structure has three parts: data set processing, model construction, and model evaluation. The data set processing contains data curation and pre-processing, data labelling, data segmentation, data normalization and balancing and data splitting. The model construction interprets the applied CNN model and the spatial pyramid pooling structure. The model evaluation includes performance metric, test and evaluation, optimization using validation data, and final prediction using test data.

The application of large CNN with many layers on a small data set will lead to over-fitting, that is, the phenomenon that the specific data set is too closely or accurately matched, so that other data cannot be well fitted, or future observations cannot be predicted. This will cause the CNN model to identify difficult laryngoscopes well on the dataset used, but it cannot generalize to identify whether other laryngoscopes are difficult. In this work, there are three convolution layers and three dense layers in this convolutional neural network. Normalization, pooling, and random dropout operations were used after each layer of CNN to prevent
overfitting. The activation function uses “Relu” in the middle module, and “Sigmoid” in output module. Loss calculated by binary cross-entropy. Optimization uses Adam. If the loss value remains unchanged within ten steps, the early stop will be executed, and the learning rate will be reduced. The resolutions of the square images were reduced to 64 × 64 pixels.

4.2. Fine-grained visual classification

The space pyramid pooling (SPP) [38] structure was applied to eliminate fixed size constraints. More precisely, the SPP [39] structure was added after final convolutive operation. It aggregated characteristics at different scales and produced the uniform-size output, that was inserted in a full-connected module (or classifier). Certain characteristics “aggregation” in the more profound model hierarchization was carried out to reduce dependence on early cropping and deformation [40].

According to previous studies and recommendations from professional anesthesiologists, the spatial information is important for recognize and classify difficult laryngoscope images. Spatial pyramid pooling can preserve space-based information through sharing in local space cells, thereby improving categorization efficiency.

4.3. Visual transfer application

Rather than being trained from the ground up, deep learning models can be tailored from available and well-trained structures like Google Net [41], ENet [42] or ImageNet. Training available structures via fine tuning hyper-parameters and weight values of tasks irrelevant to current target, known as transfer learning (TL), allows for the development of efficient and high-performance structures [43]. In this study, three transfer learning methods (DenseNet-121, ResNet-50 [44], and VGG-16 [45]) are utilized to extract characteristics from laryngoscope data. They are previously trained on ImageNet. The resolutions of the square images of the training and validation sets were reduced to 224 × 224 pixels. Figure 7 shows the SPP architecture and DenseNet-121. In the figure, SPP contains three-scale pooling structures, 16 × 256, 4 × 256, and 1 × 256; DenseNet-121 contains four dense blocks.

Several traditional machine learning algorithms were used in this paper, they are Random Forests, Logistic Regression, SVM, Naïve Bayes, KNN, XGBoost, as well as cosine similarity algorithm. Resolutions of the square images of the training and validation sets were reduced to 180 × 180 pixels. All laryngoscope X-rays were converted to histogram, and then applied histogram equalization.
5. Results

5.1. DL-based models

The results of CNN and three transfer learning models: DenseNet-121, VGG-16, and ResNet-50. The dataset used was augmented. The outcomes (Table 1) were the average outcome for 10 model executions. Table 1 indicates the outcomes of TF structures were not improved (Figure 8).

Table 2 shows the results of CNN, CNN applying spatial pooling, and DenseNet-121. The dataset used was pre-processed by data segmentation and data labelling. In this part, different labels were applied. Label-123 means all three labels were used on the dataset. Label-1 means only
the first indicator was applied on the dataset, which is the vertical gap between the hyoid’s highest point and mandibular body. Label-2 means only one variable, the atlanto-occipital gap, was applied on the dataset. Label-3 means only one variable, the corner between lower margin of the second and the sixth vertebrae cervicale in the neutral position, was applied on the dataset. Different labels were applied on the dataset to obtain the rank of which variable is the most related factor with difficult laryngoscopy.

The results shows that the spatial pyramid pooling can slightly improve the result of CNN before pre-processing. Data augmentation cannot improve the test accuracy. The result of Label-123 is the best, then is the Label-3, which is the corner between the second vertebrae cervicale’s lower margin and the sixth vertebrae cervicale’s lower margin at neutral condition, is slightly better.

5.2. ML-based models

The result of CNN applying spatial pyramid pooling, and traditional ML structures: Logistic Regression (RF), SVM, XGBoost, Random Forests (RF), Naïve Bayes (NB), KNN, as well as Cosine similarity. The dataset used were labelled by all three labels, without data segmentation. The dataset structures: Logistic Regression (RF), SVM, XGBoost, Random Forests (RF), Naïve Bayes (NB), KNN, as well as Cosine similarity. The dataset used were labelled by all three labels, without data segmentation. The results of CNN and Transfer learning are the average result of 10 executions.

Table 1 shows that some ML models have better test accuracy than DL-based models. The Naïve Bayes model has the best result, which is 86.62%. The next best performer is Logistic Regression, with an accuracy of 86.56% on the test set (Figure 9).

6. Discussion

Adequate management of the difficult laryngoscopy is one of the most important challenges for anesthesiologists during their daily clinical practice [11]. However, there is no guideline and golden standard for difficult laryngoscopy detection. There are not many studies on the application of preoperative X-ray data to distinguish patients with difficult laryngoscopy. Label-1, in our research, defined as the vertical gap between the hyoid’s highest point and mandibulae, reflects epiglottis’ position and it was already explained via Naguib [33] and Chou [34]. Naguib suggested that Label-1 did not differ in challenging and easy laryngoscopic cluster. While Chou demonstrated that Label-1 was lengthier in challenging laryngoscopic category compared to simple category. In this article, Label-1 was considered as a very important predictor of difficult laryngoscopy which was consistent with Horton’s findings [35] demonstrating that the gap between mandibulae and hyoid was consistently about 50% with respect to the gap between glottis and mandible. A big gap between mandibular body and hyoid’s highest point indicates a profound glottis. As a result, it is difficult for the anesthesiologist to expose the whole glottis because tissue places in front of the vocal cords. The test accuracy for Label-1 was 83.06%, illustrating its good prediction accuracy. Label-2 is atlanto-occipital interval, which is a gap from the first cervical vertebra to occipital bone in unbiased intubated participants. Impaired atlanto-occipital interval causes an elevated incidence of difficulties at laryngoscopy [36]. Label-2 is associated with the occipito-atlanto complex as well as mandibular protrusion. The prevalence of challenging laryngoscopy in patients with occipito-atlas complex lesions is higher than that in patients with suboccipito-atlas complex lesions [36]. Besides, shorter Label-2 distance may reflect reduced movement amplitude and mild amalgamation of the atlanto-occipital articulation. The atlantooccipital space was highly effective in the Assistant technique cluster as well as in the Macintosh laryngoscopy cluster. The test accuracy of Label-2 was 83.20%, representing high accuracy. In this study, the Label-3 is the corner between the second vertebrae cervicale’s inferior edge and the sixth vertebrae cervicale’s bottom edge in neutral place. Label-3 indicated the limited flexion of cervical spines, that may lead to challenging laryngoscopy. In this case, predictors mirroring cervicale vertebra movability might be more predictive. To conclude, Label-3 was an invaluable factor for challenging laryngoscopy prediction with test accuracy 83.33%.

On a dataset with a total of 671 images, the performance of difficult laryngoscopy decision under single index and comprehensive multiple indicators was verified uniformly for the first time. In this study, the prediction accuracy of the three indicators of Label-1, Label-2, and Label-
3 is similar (83.06% vs. 83.20% vs 83.33%). When the three indicators acted together, the performance was 83.74%. The effectiveness of the three indicators is not significantly different individually and comprehensively, Label-3 was slightly better than others when compared separately, suggesting that anesthesiologists can flexibly choose appropriate measurement indicators according to the actual situation to predict difficult laryngoscopy in clinical trials. If some indicators are blocked by neck objects, consider replacing other indicators.

Early work has been largely based on the development and application of DL-based methods to assist medical image analysis. This study compared the performance of DL with ML methods to aid detection and classification for difficult laryngoscopy. It provides an insight into the recent application of AI/ML for the classification of difficult laryngoscopy using preoperative cervical spine X-ray images. The results have shown that traditional machine learning methods outperformed commonly used convolutional neural network on medical image classification and potentially for additional image-driven medical diagnostics. Some traditional ML models achieved better test accuracy than DL-based models, like LR and Naïve Bayes. For a small and complex laryngoscopy X-ray dataset, ML can achieve remarkable accuracy and precision in forecasting data, that can exceed the ability of standard statistical techniques and human judgement. Interestingly, transfer learning models do not always improve accuracy. DenseNet-121 can achieve better accuracy. However, VGG-16 and ResNet-50 did not improve the results of difficult laryngoscopy. The pre-trained dataset of all these transfer learning models is ImageNet, which is very different from laryngoscope X-ray dataset. Available open-source medical datasets, such as MedMnist [46], have not been pre-trained on transfer learning models.

Discriminative CNNs can boost the performance of image classification, by imposing a metric learning regularization [47] or SPP structures on functional learning. The SPP structure was applied and slightly improved the result of CNN before pre-processing. On labelled dataset, CNN applied spatial pyramid pooling behaved worse than CNN, which may because SPP extracted too many features for the small 3-layer CNN. Findings demonstrate SPP and spatial information help improve the performance of difficult laryngoscopy classification but need larger dataset. Attention–based algorithms have the potential to be a powerful tool to detect difficult airway.

Data segmentation did not improve the result. The gradient-based watershed segmentation structure was introduced for segmenting medical pictures in the spatial domain. For the complicated dataset, DL-based

---

**Table 2. The results of CNN and DenseNet-121 applying pre-processing. SSP is spatial pyramid pooling.**

| Models          | Data segmentation | Data labelling  |
|-----------------|------------------|----------------|
|                 | Before pre-processing | After pre-processing |
| CNN             | 80.62%           | 80.62%         |
| CNN + SSP       | 80.77%           | 80.62%         |
| DenseNet-121    | 80.95%           | 80.77%         |

3 is similar (83.06% vs. 83.20% vs 83.33%). When the three indicators acted together, the performance was 83.74%. The effectiveness of the three indicators is not significantly different individually and comprehensively, Label-3 was slightly better than others when compared separately, suggesting that anesthesiologists can flexibly choose appropriate measurement indicators according to the actual situation to predict difficult laryngoscopy in clinical trials. If some indicators are blocked by neck objects, consider replacing other indicators.

Early work has been largely based on the development and application of DL-based methods to assist medical image analysis. This study compared the performance of DL with ML methods to aid detection and classification for difficult laryngoscopy. It provides an insight into the recent application of AI/ML for the classification of difficult laryngoscopy using preoperative cervical spine X-ray images. The results have shown that traditional machine learning methods outperformed commonly used convolutional neural network on medical image classification and potentially for additional image-driven medical diagnostics. Some traditional ML models achieved better test accuracy than DL-based models, like LR and Naïve Bayes. For a small and complex laryngoscopy X-ray dataset, ML can achieve remarkable accuracy and precision in forecasting data, that can exceed the ability of standard statistical techniques and human judgement. Interestingly, transfer learning models do not always improve accuracy. DenseNet-121 can achieve better accuracy. However, VGG-16 and ResNet-50 did not improve the results of difficult laryngoscopy. The pre-trained dataset of all these transfer learning models is ImageNet, which is very different from laryngoscope X-ray dataset. Available open-source medical datasets, such as MedMnist [46], have not been pre-trained on transfer learning models.

Discriminative CNNs can boost the performance of image classification, by imposing a metric learning regularization [47] or SPP structures on functional learning. The SPP structure was applied and slightly improved the result of CNN before pre-processing. On labelled dataset, CNN applied spatial pyramid pooling behaved worse than CNN, which may because SPP extracted too many features for the small 3-layer CNN. Findings demonstrate SPP and spatial information help improve the performance of difficult laryngoscopy classification but need larger dataset. Attention–based algorithms have the potential to be a powerful tool to detect difficult airway.

Data segmentation did not improve the result. The gradient-based watershed segmentation structure was introduced for segmenting medical pictures in the spatial domain. For the complicated dataset, DL-based

---

**Table 3. The results of CNN and traditional ML models.**

| Models         | Test accuracy | F1 Score  | Average precision score |
|----------------|---------------|-----------|-------------------------|
| CNN            | 83.74%        | 0.6627    | 0.8317                  |
| LR             | 86.56%        | 0.9868    | 0.8293                  |
| SVM            | 84.62%        | 0.8818    | 0.8359                  |
| XGBoost        | 84.37%        | 0.9074    | 0.8373                  |
| RF             | 84.21%        | 0.908     | 0.832                   |
| NB             | 86.62%        | 0.9075    | 0.8373                  |
| KNN            | 84.13%        | 0.9082    | 0.832                   |
| Cosine similarity | 84.13%      | 0.9083    | 0.832                   |
segmentation without discarding other information may be more useful. Using this type of dataset, which has complex features, labelling is an effective way to achieve meaningful results. The order of importance of the three predictors was proposed to aid anaesthesiologists to recognise difficult laryngoscopy. Label-1 is associated with difficult laryngoscopy, and the result is consistent with previous study [4]. Label-2 is associated with occipito-atlanto complex and is related to mandibular protrusion [6]. The incidence of airway difficulties in patients with occipito-atlanto complex lesions was higher than that in patients with disease below the complex [48]. Label-3 achieved the best performance, and might represent the best indicator of challenging laryngoscopy reflecting cervical movability, consistent with previous result [5]. The results demonstrated that label-3 was slightly better than label-2, and both were slightly better than label-1. Our results are similar to the study from Liu et al. [6], label-3 is superior to label-1, but [6] they did not compare the relationship between label-3 and label-2 or that between label-2 and label-1.

There are certain limits, though, in this work. Firstly, the data in the research is preoperative X-ray pictures of participants with cervical spondylosis, and a multicentre study may make the results of the study more convincing. Secondly, this research is a prospective study, in order to predict the assistive technology application’s necessity and possibility during intubation, which may provide more reference value for clinical anaesthesia. In addition, there may be some measurement related errors. Future work will focus on applying multimodal databases in classifying difficult laryngoscopy and seeking to apply pre-trained medical related datasets on transfer learning models.

Declarations

Author contribution statement

Xiaoxiao Liu; Launcelot McGrath: Conceived and designed the experiments; performed the experiments; Wrote the paper.

Colin Flanagan; Yiming Lei, Jiangzhen Guo: analysed and interpreted the data.

Jingchao Fang; Jun Wang and Xiangyang Guo, Harry McGrath, Yongzheng Han: Contributed reagents, materials, analysis tools or data.

Funding statement

This work was supported by Young Scholar Research Grant of Chinese Anesthesiologist Association (21900007), Key Clinical Projects of Peking University Third Hospital (BYSVDZ2021013), Beijing Haiyan District Innovation and transformation project (HDCHZHBR2021202) and Clinical Medicine Plus X - Young Scholars Project, Peking University, The Fundamental Research Funds for the Central Universities PKU2022LCXQ031.

Data availability statement

The data that has been used is confidential.

Declaration of interest’s statement

The authors declare no conflict of interest.

Additional information

No additional information is available for this paper.

References

[1] H.J. Kim, et al., Anterior neck soft tissue measurements on computed tomography to predict difficult laryngoscopy: a retrospective study, Sci. Rep. 11 (1) (2021) 8438.

[2] L. Li, et al., Airtraq laryngoscope: a solution for difficult laryngeal exposure in phonmicrosurgery, Acta Otolaryngol. 137 (6) (2017) 635-639.

[3] H. Gonzalez, et al., The importance of increased neck circumference to intubation difficulties in obese patients, Anesth. Analg. 106 (4) (2008) 1132–1136, table of Table 3.

[4] Y.Z. Han, et al., Radiologic indicators for prediction of difficult laryngoscopy in patients with cervical spondylosis, Acta Anaesthesiol. Scand. 62 (4) (2018) 474–482.

[5] Y. Zhou, et al., Preoperative X-ray C2C6AR is applicable for prediction of difficult laryngoscopy in patients with cervical spondylosis, BMC Anesthesiol. 21 (1) (2021) 111, 20 (1) (2020) 238.

[6] A.esta, et al., Deep learning-enabled medical computer vision, NPJ Digit. Med. 4 (1) (2021) 5.

[7] X. Kuang, et al., Accurate and rapid prediction of tuberculosis drug resistance from genome sequence data using traditional machine learning algorithms and CNN, Sci. Rep. 12 (1) (2022) 2427.

[8] Y.Z. Han, et al., Neck circumference to inter-incisor gap ratio: a new predictor of difficult laryngoscopy in cervical spondylosis patients, BMC Anesthesiol. 17 (1) (2017) 55.

[9] M. Avanzo, et al., Artificial intelligence applications in medical imaging: a review of the medical physics research in Italy, Phys. Med. 83 (2021) 221–241.

[10] H. Zhang, et al., IR-net: low-rank spatial-spectral network for hyperspectral image denoising, IEEE Trans. Image Process. 30 (2021) 8743-8758.

[11] M.A. Morid, A. Borjali, G. Del Fiol, A scoping review of transfer learning research on medical image analysis using ImageNet, Comput. Biol. Med. 128 (2021), 104115.

[12] D.S.W. Ting, et al., AI for medical imaging goes deep, Nat. Med. 24 (5) (2018) 539–540.

[13] C. Yang, et al., Pre-treatment ADC image-based random forest classifier for identifying resistant renal adenocarcinoma to neoadjuvant chemoradiotherapy, J. Color. Res. 35 (1) (2020) 101–107.

[14] N. Jia, et al., Subclinical diabetic peripheral vascular disease and epidemiology using logistic regression mathematical model and medical image registration algorithm, J. Healthc. Eng. 2022 (2022), 2116224.

[15] C.V. Chang, et al., SVM-enabled intelligent genetic algorithmic model for realizing efficient universal feature selection feature in breast cyst lesion acquired via ultrasound sensing systems, Sensors (Basel) 20 (2) (2020).

[16] H. McGrath, et al., Future of artificial intelligence in aesthetics and pain management, J. Biosci. Med. 7 (11) (2019) 111–118.

[17] N.C. D'Amico, et al., Radiomics-based prediction of overall survival in lung cancer using different volumes-of-interest, Appl. Sci. 10 (18) (2020).

[18] E. Lella, et al., Machine learning and DWT brain communicability networks for Alzheimer’s disease’s detection, Appl. Sci. 10 (3) (2020).

[19] M. Avanzo, et al., Electron density and biologically effective dose (BED) radiomics-based machine learning models to predict late radiation-induced subcutaneous fibrosis, Front. Oncol. 10 (2020) 496.

[20] N. Amoroso, et al., Deep learning and multiplex networks for accurate modeling of brain age, Front. Aging Neurosci. 11 (2019) 115.

[21] A. Lombardi, et al., Extensive evaluation of morphological statistical harmonization for brain age prediction, Brain Sci. 10 (6) (2020).

[22] J.P. Kandhasamy, et al., Diagnosis of diabetic retinopathy using multi level set segmentation algorithm with feature extraction using SVM with selective features, Multimed. Tool. Appl. 79 (15–16) (2019) 10581–10596.

[23] S. Saha, et al., Feature selection for facial emotion recognition using cosine similarity-based harmony search algorithm, Appl. Sci. 10 (8) (2020).

[24] Y. Lei, et al., Subclinical diabetic peripheral vascular disease detection, Appl. Sci. 10 (3) (2020).

[25] C. Frerk, et al., Difficult Airway Society 2015 guidelines for management of anticipated difficult intubation in adults, Br. J. Anaesth. 115 (6) (2015) 827–848.

[26] E.D. Cubuk, et al., AutoAugment: Learning Augmentation Policies from Data, 2019.

[27] F. Liu, et al., ACPL: Anti-curriculum Pseudo-labeling for Semi-supervised Medical Image Classification, 2021.

[28] S. Huang, L. Shen, M. Lungren, GluRA: A Multimodal Global-Local Representation Learning Framework for Label-Efficient Medical Image Recognition, 2021.

[29] X. Hu, et al., Semi-supervised Contrastive Learning for Label-Efficient Medical Image Segmentation, 2021.

[30] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet Classiﬁcation, 2012.

[31] X. Hu, et al., Semi-supervised Contrastive Learning for Label-Efficient Medical Image Recognition, 2021.

[32] U.K. Acharya, S. Kumar, Genetic algorithm based adaptive histogram equalization (GAAHE) technique for medical image enhancement, Optik (2021) 230.

[33] N. Jia, et al., Subclinical diabetic peripheral vascular disease and epidemiology using logistic regression mathematical model and medical image registration algorithm, J. Healthc. Eng. 2022 (2022), 2116224.

[34] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet Classiﬁcation, 2012.
[41] C. Spampinato, et al., Deep learning for automated skeletal bone age assessment in X-ray images, Med. Image Anal. 36 (2017) 41–51.

[42] S. Moccia, et al., Development and testing of a deep learning-based strategy for scar segmentation on CMR-LGE images, MAGMA 32 (2) (2019) 187–195.

[43] S. Basaia, et al., Automated classification of Alzheimer's disease and mild cognitive impairment using a single MRI and deep neural networks, Neuroimage Clin. 21 (2019), 101645.

[44] S. Targ, D. Almeida, K. Lyman, Resnet in Resnet: Generalizing Residual Architectures, 2016.

[45] K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, 2014.

[46] J. Yang, R. Shi, B. Ni, Medmnist Classification Decathlon: A Lightweight AutoML Benchmark for Medical Image Analysis, 2021.

[47] G. Cheng, et al., When deep learning meets metric learning: remote sensing image scene classification via learning discriminative CNNs, IEEE Trans. Geosci. Rem. Sens. 56 (5) (2018) 2811–2821.

[48] I. Calder, J. Calder, H.A. Crockard, Difficult direct laryngoscopy in patients with cervical spine disease, Anaesthesia 50 (9) (1995) 756–763.