Computers in Biology and Medicine 136 (2021) 104737

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Identification of difficult to intubate patients from frontal face images using an ensemble of deep learning models

Thomas E. Tavolara [a,*], Metin N. Gurcan [a], Scott Segal [b], M.K.K. Niazi [a]

[a] Center for Biomedical Informatics, Wake Forest School of Medicine, Winston-Salem, NC, USA
[b] Dept. of Anesthesiology, Wake Forest School of Medicine, Winston-Salem, NC, USA

ARTICLE INFO

Keywords:
Endotracheal intubation
Deep learning
Machine learning
Airway management
Image analysis

ABSTRACT

Failure to identify difficult intubation is the leading cause of anesthesia-related death and morbidity. Despite preoperative airway assessment, 75–93% of difficult intubations are unanticipated, and airway examination methods underperform, with sensitivities of 20–62% and specificities of 82–97%. To overcome these impediments, we aim to develop a deep learning model to identify difficult to intubate patients using frontal face images. We proposed an ensemble of convolutional neural networks which leverages a database of celebrity facial images to learn robust features of multiple face regions. This ensemble extracts features from patient images (n = 152) which are subsequently classified by a respective ensemble of attention-based multiple instance learning models. Through majority voting, a patient is classified as difficult or easy to intubate. Whereas two conventional bedside tests resulted in AUCs of 0.6042 and 0.4661, the proposed method resulted in an AUC of 0.7105 using a cohort of 76 difficult and 76 easy to intubate patients. Generic features yielded AUCs of 0.4654–0.6278. The proposed model can operate at high sensitivity and low specificity (0.9079 and 0.4474) or low sensitivity and high specificity (0.3684 and 0.9605). The proposed ensembled model outperforms conventional bedside tests and generic features. Side facial images may improve the performance of the proposed model. The proposed method significantly surpasses conventional bedside tests and deep learning methods. We expect our model will play an important role in developing deep learning methods where frontal face features play an important role.

1. Introduction

Failure of successful airway management continues to be the leading cause of anesthesia-related death and severe morbidity [1,2]. Considered the worldwide standard of care [3,4], preoperative airway assessment serves to determine the degree of difficulty with various airway management strategies [5]. This typically includes two examinations – the Mallampati test (MP) [6] and the TMD (thyromental distance) [7]. In [3], it was shown that these do not reliably predict difficult intubation – MP class $\geq$ 3 or TMD $\leq$ 3 fingerbreadths resulted in sensitivity and specificity of 32% and 85%, respectively, consistent with the historical performance [8]. Other airway examination algorithms perform modestly, with sensitivities of 20–62%, specificities of 82–97% [9,10] To overcome these impediments, anesthesiologists need better tools to predict difficult intubation to minimize treatment-related complications and healthcare expense.

Recent studies [2–4,11,12] have sought such ends through computerized analysis of facial images. Connor et al. [2] utilized FaceGen, which generates a 3D model of a patient's head using front and side face images. Sixty-one face proportions were computed using this 3D model and utilized as features in a feature selection model followed by logistic regression to identify difficult to intubate patients. However, Connor et al. utilized a relatively small cohort of patients and performed feature selection on the whole dataset rather than on a training cohort, leading to potential biases. Cuendet et al. [12] utilized automatically detected fiducial landmarks of the front and side view face images along with principal component analysis of textures of the inside of the mouth. Likewise, feature selection was performed, and a cascade of random forest classifiers to predict intubation difficulty. The method proposed by Cuendet et al. overcomes these shortcomings of Connor et al. with a large cohort of patients (n > 900) but presumably would not generalize to the bedside, as their imaging protocol was extremely controlled and strict. To the best of our knowledge, there still lacks a comprehensive automated image analysis model to identify difficult intubation from

* Corresponding author. Wake Forest School of Medicine, 486 Patternson Avenue Winston-Salem, NC, 27101, USA.
E-mail address: ttavolar@wakehealth.edu (T.E. Tavolara).

https://doi.org/10.1016/j.compbiomed.2021.104737
Received 6 May 2021; Received in revised form 1 August 2021; Accepted 2 August 2021
Available online 4 August 2021
0010-4825/© 2021 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

frontal and profile view facial images in the wild.

Inspired by these studies, advances in deep learning [13], and a recent study utilizing frontal face images to predict genetic disorders [14], we seek a deep learning model based on the analysis of facial images superior to conventional bedside tests to ultimately improve airway management and patient safety. Popular deep learning methods make heavy use of transfer learning [15] and data augmentation. This out-of-the-box approach co-opts deep learning models trained on millions of images of commonplace objects [16] and retrains a subset of model parameters on a target domain with a small number of images [17,18]. In this study, we employ these techniques as a baseline model. Furthermore, we compare with a deep learning model pretrained solely on front face images. Finally, we demonstrate that our own method based on deep multiple instance learning outperforms these conventional deep learning methods and outperforms conventional bedside MP and TMD tests on a cohort of front face images of patients taken in the wild.

2. Materials and methods

We propose a multi-stage ensembled deep learning model to identify difficult to intubate patients from front face images. Briefly, our methodology utilizes a large database of celebrity facial images (CASIA-Webface [19]) to train 11 convolutional neural networks (CNN) on 11 facial regions. These 11 models are used as feature extractors on patient images and are subsequently classified by a set of respective 11 attention-based multiple instance learning (MIL) [20] models. Then, through majority voting, the patient is classified as difficult or easy to intubate. We hypothesize that anesthesiologists' visual assessment and features beyond human comprehension can be modeled through deep learning to identify difficult to intubate patients.

2.1. CASIA-webface

Typically, CNNs are trained on hundreds of thousands to millions of images to generalize to unseen data. Unfortunately, few datasets reach these magnitudes with the exception of those like ImageNet [16], consisting of everyday objects such as trees, cars, and people, and others like it. However, the features learned on such objects may not be useful for transfer learning. Therefore, we sought to train a custom CNN using a face database.

CASIA-Webface [19] is a large dataset of celebrity faces (Fig. 1). It contains 494,414 images from 10,575 different subjects that were obtained through a semi-automated process of detecting and clustering celebrity images on IMDb. All images are 250 × 250 pixels. Each face in the dataset was subject to landmark detection, face alignment, and region extraction as described in the following sections.

2.2. Landmark detection

Landmark detection is the process of automatically placing points on an image that are consistent across similar images. On a face image, for example, we would like to be able to consistently place a landmark for an eye across many different face images. The motivation for landmark detection is three-fold. First, CASIA-Webface contains faces that may not be present or may be obstructed, and so these images should be ignored. Second, the subsequent step of face alignment requires landmarks. Third, our proposed method requires extracting specific face regions, and thus landmarks are required.

Landmark detection was performed with dlib [21] and a method developed by Kazemi et al. [22]. These libraries take a given input face image and output 68 landmarks on the face, given that a face is detected in the image and all landmarks are visible. The method is robust to pose, meaning that landmarks can be placed on face images taken at an angle (Fig. 1). Briefly, dlib's face detector utilizes HOG [23] features extracted from various input image scales using a sliding window scheme. These features from different regions of the face are then classified using and SVM into part of a face or part of the background. Therefore, this first step *localizes* the face. Using this localized face, the method by Kazemi et al. [22] utilizes forest of gradient boosted regression trees trained on raw pixels values to detect various landmarks around the face. Each tree is trained on the residuals of the previous tree in the ensemble, and therefore improves its estimation of landmarks iteratively. These landmarks include points along the perimeter of the eyes, nose, mouse, lips, chin, cheeks, and eyebrows (Fig. 1). A custom neck landmark was estimated by computing the distance between 1) the midpoint of the line formed between the eyes and 2) the middle of the chin, halving it, and taking the point along the line formed by 1) and 2) below the chin.

2.3. Face alignment

The intuition behind face alignment transformation is that when training convolutional neural networks on face images, one would ideally want the locations of various landmarks to be approximately in the same place relative to one another. For example, the line formed between the eyes should be parallel to the bottom perimeter of the image or that the nose should be in the exact center of the image. This way, a model need not be translation- or scale-invariant. This reduces the number of parameters that a CNN needs to learn.

Face alignment was performed using OpenFace [24]. This process utilizes a set of detected landmarks of an input face, derived from the previous step in the overall method. OpenFace then applies an affine transformation with six degrees of freedom on a subset of these points (i.e. their x-y coordinates in the image) to be as close as possible to a template face (i.e., "average" face) with the same landmarks. The

Fig. 1. Landmark detection on example images of celebrities, from pexels.com and unsplash.com, freely licensed with no permission. From left to right – Martin Luther King Jr., Marylin Monroe, Barack Obama, and Audrey Hepburn.

transformation matrix includes translation, scaling, and rotation terms thus acts to align *and* rescale the face. This preserves points, straight lines, and planes on the 2D image. Three sets of points were utilized for face alignment, creating three distinct versions of the CASIA-Webface dataset (Fig. 2). The first utilized points on the outer corner of the eyes and bottom of the nose; the second utilized the inner corners of the eyes and bottom center of the lips; and the third utilized no alignment. The former two were fit to the same "average" face default template in OpenFace.

2.4. Face region extraction

Finally, region extraction was performed by cropping 100 × 100 images centered on specific landmarks (Fig. 2). The motivation was to allow each subsequent CNN-based feature extractor (next section) to focus on one region of the face. Furthermore, this allows ensembling of many regions of the face and allows a degree of interpretability (i.e., which region of the face is most predictive of difficult intubation). These landmarks included left and right eyes, nose bridge, nose, mouth, chin, left and right cheek, left and right jaw, and neck (total = 11) as in Fig. 1. Each image was converted to grayscale, as we had no reason to believe that color would influence the difficulty of intubation.

2.5. CNN-based face region feature extractor

The motivation behind feature extraction is two-fold. First, extracted features reduce the dimensionality of data. In the proposed method, 100 × 100 pixel images (10,000 pixels total) are each mapped to a 320-dimensional feature space. Second, extracted features are abstracted representations of their respective inputs (in our case, face regions), which are often imperceivable by the human eye. The underlying assumption is that if a CNN can be trained to somewhat accurately classify celebrities based only on a single face region, then the learned face region features (Fig. 2) must be robust to facial features.

After the pre-processing steps described above, a CNN was trained for each version of the CASIA-Webface dataset (from different face alignments as in Fig. 2) to classify celebrities based on a single face region. The dataset was divided into an 85/5/10 ratio for training, validation, and testing as in [14]. The model architecture is depicted in Fig. 3. Weights were initialized using He [25].

Each CNN was trained in two phases. The first phase was at an initial learning rate of 0.001, for 40 epochs, and with a mini-batch size of 128. The second phase was at an initial learning rate of 0.0001 with a momentum of 0.9 for ten epochs, and with a mini-batch size of 128. These two phases enable exploration and exploitation of the cost function – the first training phase explores the parameter space more broadly while the second phase fine-tunes to a local minimum. In total, 33 distinct CNNs were trained (3 face alignments, 11 face regions). These custom CASIA-Webface trained face region feature extractors ("Face Region Feature Extractor", FRFE) were utilized to extract features from patient face regions.

2.6. Patient dataset

For each patient subject, a frontal face image along with MP and TMD were collected by an experienced anesthesiologist pre-operatively. Images were captured at the bedside of the patient with study approval from Wake Forest University Institutional Review Boards (IRB00036442) and patient consent. Each patient had a "*ground truth*" label of 0 (easy) or 1 (hard) based on the difficulty of intubation during general anesthesia. Patients were defined as easy to intubate if only a single attempt with a Macintosh 3 blade was needed, resulting in a grade 1 laryngoscopic view. Difficult intubation was defined by at least 1 of the following – more than one attempt by an operator with at least 1 year of anesthesia experience, grade 3 or 4 laryngoscopic view on a 4-point scale, need for a second operator, or nonelective use of an alternative airway device such as a bougie, fiberoptic bronchoscope, or intubating laryngeal mask airway [2]. In total, 76 patients were difficult to intubate, and 429 were easy to intubate. Prior to the landmark detection, face alignment, face region extraction steps described above, a data augmentation step was taken in which each patient image was scaled using a factor of 0.75–1.1 in steps of 0.05 and rotated between −5 and 5° in steps of 1. In total, there were 88 different scale and rotation combinations for each patient front face image. Each resulting patient face region was subject to feature extraction by their respective FRFE model (Fig. 2a). All difficult and a random subset of 76 easy to intubate patients were selected for model cross-validation given severe class imbalance. The remaining easy to intubate patients were discarded.

2.7. Attention-based MIL

The motivation behind utilizing MIL derives from the data augmentation step. In generating multiple scales and rotations of the same face region, how to combine all these augmentations into a single

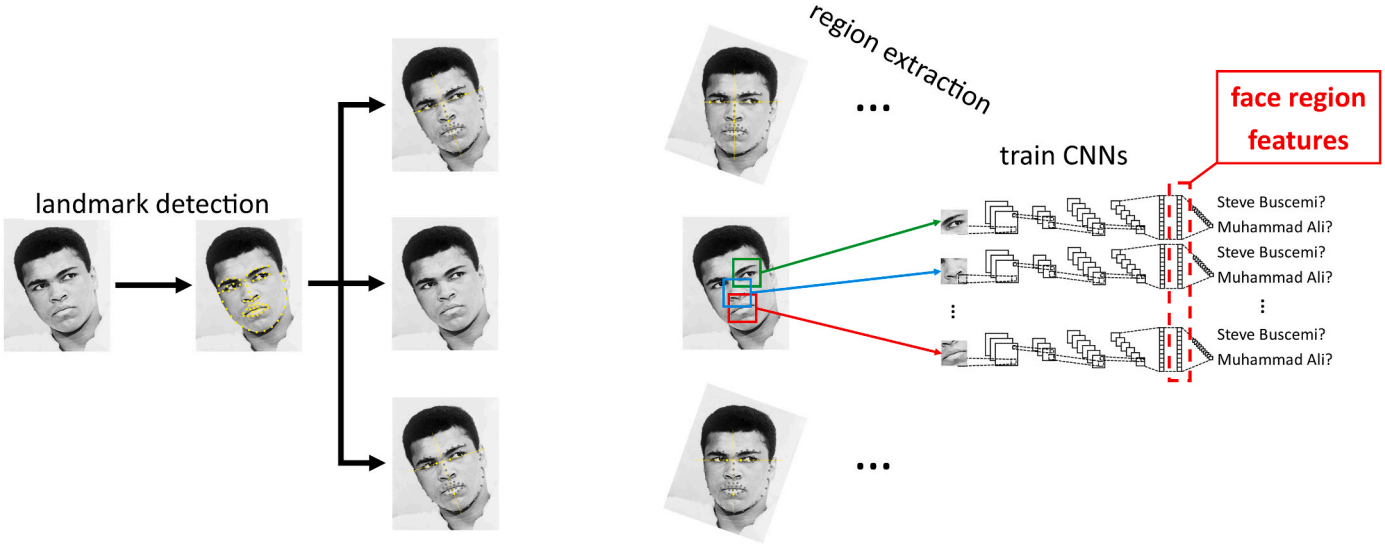

Fig. 2. Landmark detection, face alignment using different sets of points (outer points of eyes and bottom of the nose; no alignment; inner points of eyes and bottom of the lip), face region extraction, and finally CNN training on CASIA-Webface. Image of Muhammad Ali, from Wikimedia Commons, public domain.

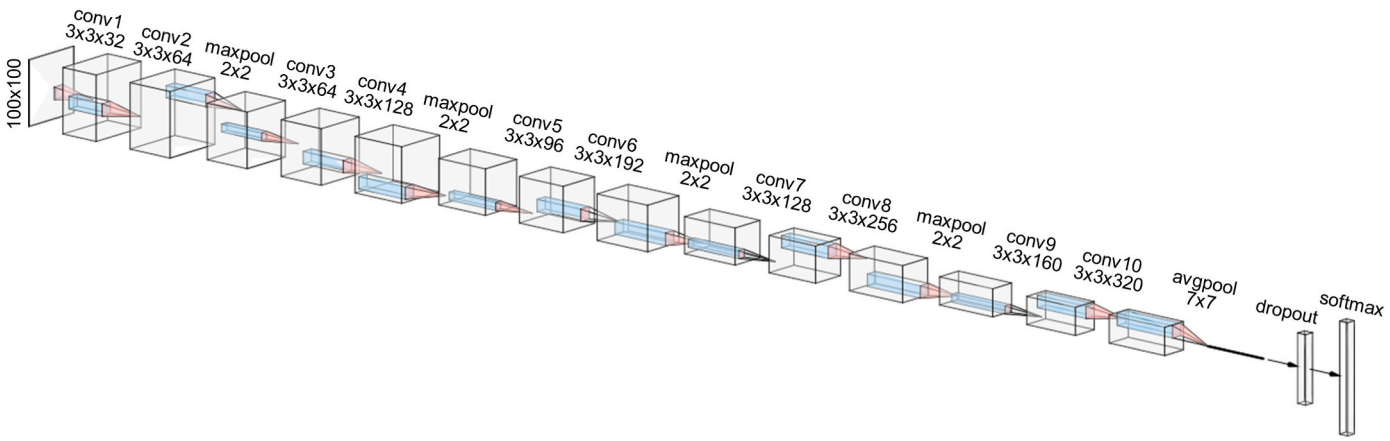

Fig. 3. CNN architecture. Convolutional layer parameters are expressed as (filter size, filter size, number of filters), all have a stride of 1, and all are followed batch normalization and ReLU. maxpool layer parameters are expressed as (size, stride). softmax is composed of a fully connected layer followed by a softmax activation function.

decision feature vector becomes problematic. Which scale/rotation is most apt for predict difficult intubation? Simply concatenating the features extracted using different augmentations for a single face region is risky, as a curse of dimensionality problem arises – the number of features far exceeds the number of samples (patients) available. This can potentially lead to overfitting. Therefore, MIL is a must. However, the function by which augmentations are aggregated then becomes the question. We address this through attention-based [20] MIL [26].

Multiple instance learning (MIL) [26] is a machine learning paradigm in which labels are assigned to collections of examples (bags) rather than the examples themselves (instances). The idea arises from situations in which explicit labels are known for collections of examples, but individual example labels are not or cannot be known, but they are implicit. Classification is thus done on bags rather than instances. This can be done through several mechanisms – for example, classifying each individual instance and aggregating their decisions or aggregating embedded instances then performing classification on the bag-level embedding.

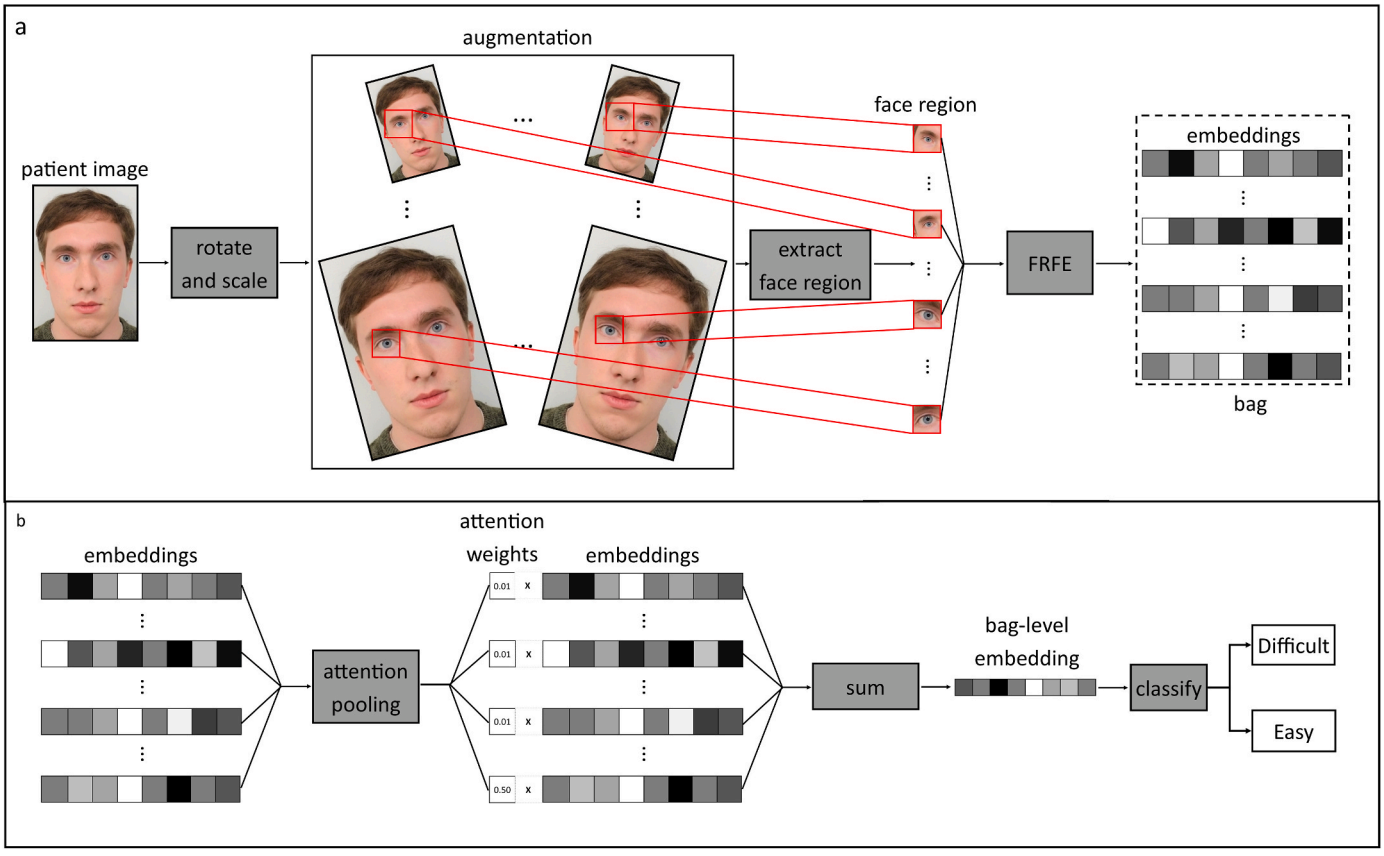

Fig. 4. Attention-based MIL bag composition and aggregation. a) Patient image is rotated and scales to achieve augmentation. A specific face region is cropped (left eye in this example), and its features are extracted by FRFE models. The resulting embeddings comprise a bag for MIL. b) Attention pooling computes weights for each embedding and then computes a weighted sum of the embeddings with respective attention weights. The subsequent bag-level embedding is classified as difficult to intubate or easy to intubate. "Patient image" is of first author.

In the case of attention-based pooling [20], an attention weight is computed for each embedded instance (i.e., features of a face region) using a two-layer convolutional neural network. The aggregation function is then a weighted sum of embedding instances using these attention weights. Then, the bag-level embedding is classified. We opted for this flavor of aggregation, as it automatically learns the aggregation function rather than us having to decide what that function should be (e.g., mean or max).

In the context of the current task, bags are created from the augmentations produced for each patient face region (Fig. 4a) and fused into a bag-level embedding using attention pooling (Fig. 4b). So, each bag contains 88 feature vectors extracted (yielded through augmentation) from the FRFE models. Therefore, for each FRFE, a distinct MIL model was trained for each face alignment and face region combination (33 models total). Each MIL model was trained with a learning rate of 5e-4 with a momentum of 0.9 for 20 epochs. Training was halted if training accuracy did not improve for five epochs.

2.8. Experimental design

We utilized two strategies to estimate the performance of our 33 FRFE models (11 face regions and three face alignment strategies). The first was to retrain the last layer of each FRFE model to predict the difficulty of intubation based only on respective face regions. In other words, the last layer of the chin FRFE model (Fig. 2) was replaced and retrained to predict difficult intubation from only images of the patient chins. This was carried out using a ten-fold cross-validation. Retraining of that last layer was carried out at a learning rate of 0.00001 with a momentum of 0.9 over 10 epochs. During validation of each fold, a classification for an individual patient's single face region was carried out by aggregating the output probabilities of each scale/rotation augmentation (i.e., 88 outputs were summed). The output with the maximum sum was deemed the overall decision for that face region of that patient. Finally, for the overall classification of a patient, a majority voting was carried out across all face region models (11 total; majority = at least 6).

The second strategy was to utilize the FRFE models as feature extractors and aggregate face region embeddings using attention-based MIL. This was carried out using a leave-one-out cross-validation. The output of each MIL model is a prediction for a face region of a patient. Therefore, majority voting was utilized to come to a consensus across all MIL face region models (11 total; majority = at least 6).

2.9. Comparison methods

We provide comparison methods to demonstrate the importance of a face-detection, data augmentation, MIL, and ultimately a domain-specific feature extractor. In our first comparison ("baseline"), we utilize Inception v4 [27] pretrained on ImageNet for transfer learning, both retraining only the last layer and fine-tuning the whole network. For the latter, a weight bias of 10x was given to the parameters in the last fully connected layer. In our second comparison ("face cropped"), an additional set of models was trained using face images which were automatically cropped for just the face using dlib (thereby removing background). In our third comparison ("face cropped augmentation"), another set of models was trained using an augmented patient dataset (as described in the Patient Dataset Section). Each model was trained using the same ten folds as described in the Experimental Design Section for a maximum of 20 epochs with a learning rate of 3×10^{-4}. Training was halted if the validation accuracy of each fold did not decrease for five epochs. Our fourth comparison utilized ImageNet pretrained Inception v4 features in the MIL paradigm (rather than features from the custom pretrained face region models. These were trained using the same experimental parameters as described in the Experimental Design Section.

Each model is assessed using sensitivity, specificity, AUC, Matthew's correlation coefficient (MCC), and F1-score. Confusion matrices are reported in the supplemental. Furthermore, statistical comparisons are made between the proposed model and 1) comparison deep learning methods and 2) MP and TMD tests using a one-sided McNemar's test [28].

3. Results

3.1. Baseline models

In our first set of experiments on the patient database, we cross-validated conventionally inspired baseline comparisons using Inceptionv4 models pretrained on ImageNet (see Table 1). These results are reported in Tables 2 and 3.

The best performing models in these conventional examples are clearly those in which all model weights are tuned (i.e., 'no freeze'). There also seems to be no difference observed when cropping the face or augmenting the dataset if only the last layer of the pretrained network is retrained. The best performing model utilized cropped faces and heavy augmentation along with fine-tuning of all model weights.

Results of MIL on ImageNet features are no better than those obtained using a single Inceptionv4 model (Table 3). There seems to be overfitting to the easy to intubate class. Results improve using an ensemble of pretrained feature extractors rather than generic features, with eleven out of twenty-seven one-sided McNemar tests yielding $p < 0.05$ when comparing the three ensembled from Table 3 to the six generic models of Table 2 and three generic ensembled models from Table 3.

3.2. Proposed models

For the proposed first strategy, simply retraining the last layer of FRFE models to predict the difficulty of intubation for patient images, the best performing model was that with alignment using the inner corners of the eye and bottom lip. It achieved a positive class accuracy (sensitivity) of 69.74% and a negative class accuracy (specificity) of 64.47% (Table 4). One noteworthy result in the case of individual face region models is the neck model with no alignment. It performs just as well as its respective ensemble model. Furthermore, at least in the cases with some alignment, ensembling seems to lead to some degree of improvement.

While these results leave room for improvement, they are much better than their MP scores and TMD clinical counterparts. Using the same cohort of patients and categorizing them using a MP score ≥ 3, the sensitivity is 35.53, specificity 78.95, and AUC 0.5748. Using a TMD ≤ 3, the sensitivity is 88.16, the specificity 3.95, and AUC 0.4933. Furthermore, comparing "ineye botlip", "outeye nose", and "no align" FRFE models to their clinical counterparts using a one-sided McNemar's test yields p-values of 8^{-22}, and 2^{-20}, and 0.0006 for TMD, respectively, and 6^{-27}, 5^{-23}, and 0.2976 for MP score, respectively.

Table 1
List of baseline and proposed models.

Inception v4 (conventional)	Inception v4 + MIL (conventional)	FRFE (proposed)	FRFE + MIL (proposed)
baseline	bridge	bridge	bridge
face cropped	chin	chin	chin
face cropped augmentation	left cheek	left cheek	left cheek
	left eye	left eye	left eye
	right jaw	right jaw	right jaw
	mouth	mouth	mouth
	neck	neck	neck
	nose	nose	nose
	right cheek	right cheek	right cheek
	right eye	right eye	right eye
	right jaw	right jaw	right jaw
	ensemble	ensemble	ensemble

Table 2
Results of retraining using conventional methods. 'baseline' refers to a retrained Inception v4 on raw images. 'face cropped' indicates a preprocessing step where faces were automatically cropped using dlib. 'face cropped augmentation' indicates preprocessing to both augment then crop the race using dlib. 'freeze' indicates that pretrained model weights were frozen except for the fully connected layer, and 'no freeze' indicates all parameters were tuned, with biases as described in the Methods. Mean and standard deviation are reported. Corresponding confusion matrices can be found in supplementary material.

		freeze	no freeze
baseline	sensitivity	63.16 ± 36.63	46.05 ± 32.27
	specificity	34.21 ± 39.59	59.21 ± 26.00
	AUC	0.4879 ± 0.1327	0.5587 ± 0.1392
	MCC	−0.0275 ± 0.24	0.0531 ± 0.22
	F1-score	55.17 ± 21.20	49.30 ± 25.52
face cropped	sensitivity	30.26 ± 36.58	60.53 ± 22.98
	specificity	67.11 ± 30.57	50.00 ± 23.48
	AUC	0.4829 ± 0.1804	0.5732 ± 0.1641
	MCC	−0.0283 ± 0.34	0.1058 ± 0.20
	F1-score	37.10 ± 28.69	58.23 ± 14.01
face cropped augmentation	sensitivity	28.95 ± 34.04	63.16 ± 22.21
	specificity	65.79 ± 35.79	57.89 ± 11.51
	AUC	0.4654 ± 0.1219	0.6278 ± 0.1530
	MCC	−0.0566 ± 0.30	0.2108 ± 0.28
	F1-score	35.48 ± 27.21	61.57 ± 16.15

In our second strategy, rather than utilizing an arbitrary aggregation function for different scales and rotations of the same region, we opted to automatically learn a function using MIL. The best performing LOO model utilized 9 of the 33 face region models, which had a leave-one-out positive class and negative class accuracy of >50%. As an ensemble, it achieved a positive class accuracy (sensitivity) of 0.7368 and a negative class accuracy (specificity) of 0.6842 with an AUC of 0.7105. This was on a different subset of easy to intubate patients, so it is not comparable to the results from the FRFE models. Table 5 summarizes the individual model results.

The MP score ≥ 3 sensitivity and specificity were 81.58 and 35.53, respectively, with AUC 0.6042. The TMD ≤ 3 sensitivity was 88.16 and specificity 10.53, with AUC 0.4661. In addition to outperforming generic feature models (Tables 2 and 3) in all nine statistical comparisons, this ensembled MIL model significantly outperformed clinical tests for difficult intubation – $p = 0.0001$ for TMD and $p = 0.0151$ for MP distance – using a one-sided McNemar's test.

In both sets of experiments, thresholds on model outputs were adjusted to achieve a sensitivity of >80%. Given this criterion, the FRFE majority voting ten-fold cross-validation sensitivity and specificity were 0.8158 and 0.4868, 0.8143 and 0.3286, and 0.8026 and 0.5263 for the "ineye botlip", "outeye nose", and "no align" experimental conditions, respectively. Similarly, the MIL majority voting cross-validation sensitivity and specificity were 0.8158 and 0.5263, respectively. These results exceed MP score in terms of sensitivity and specificity.

4. Discussion

It is interesting to note some similarities between our two proposed methods. First is mirroring. Alignment using the inner corner of the eyes and bottom lip + no alignment performed better than alignment with the outer corners of the eyes and nose (Table 4). Similarly, the face regions using the former two alignment strategies performed better in the MIL framework than the latter alignment strategy (Table 5) in the purely FRFE ensemble approach while being selected more often than alignment using outer corners of the eyes and chin in the proposed MIL method. This suggests that utilizing the outer corners of the eyes and nose as alignment targets is not as promising an approach. This may be because these landmarks are not as accurately detected as other alignment landmarks or because there is more variation in the relative position of these landmarks relative to other alignment landmarks. Second, both proposed methods had higher sensitivity than specificity. This suggests that we may utilize a larger proportion of our easy to intubate patients while developing our models. Third, both proposed methods seem to suggest that ensembling improves overall performance. This suggests that future work should continue to focus on ensembling models corresponding to different regions of the face.

In addition, we have evidence to suggest that custom feature extractors for specific facial regions also contributed to the overall performance of our models. In comparison experiments, ImageNet [16] pre-trained Inceptionv4 [27] models were utilized as either feature extractors or for fine-tuning (Tables 2 and 3) and never exceeded 60% accuracy for any fold. These results suggest that general imaging features (i.e., those learned in another domain) do not generalize well for face images and that data augmentation may not be sufficient to enhance such as small dataset. The problem with such approaches is that sometimes, these features are not sufficient to accurately adapt to an unrelated domain (such as medical images). Therefore, in this study, we pretrained a custom network on a database of only facial images to build a robust facial feature extractor (FRFE). Each of our initial set of FRFE models was able to identify celebrities with 60–70% accuracy on the testing set (true positives and true negatives divided by the total). Though this accuracy seems abysmal, it is worth noting that there were 10575 different celebrities and each model was only able to see a single face region. For example, using just the right eye, the eye model was able to achieve 70% accuracy in identifying celebrities – an attribute reflected in other FRFE models. This is remarkable, as it provides substantial support that the facial features learned by each FRFE model are robust.

Though our results suggest that our approach is superior to conventional bedside tests, there are limitations of the proposed model. First, one possible source of error could be the lack of preprocessing steps for training of the FRFEs. Unlike the patient dataset, the celebrity dataset did not undergo augmentation. This means that the variation in scale and pose of the celebrity dataset was less than the patient dataset (which was purposely augmented). In future work, we will augment CASIA-Webface like the patient dataset as well as experiment with single-scale patient images. Second, upon visual inspection of patient images, face alignment clearly fails to make faces approximately the same scale. This is because though current affine transformation can fix certain landmarks relative to the template, anatomical landmarks themselves may have considerable variation. People have different-sized eyes, noses, mouths, and neck lengths. Therefore, it may behoove us to

Table 3
Results of baseline retraining using an MIL model trained on Inceptionv4 features. 'ineye botlip' refers to alignment performed using the inner eye corners and bottom lip; 'outeye nose' refers to alignment performed using the outer eye corners and nose; and 'no align' refers to no alignment of the patient dataset. Mean and standard deviation are reported. Corresponding confusion matrices can be found in supplementary material.

		bridge	chin	left cheek	left eye	left jaw	mouth	neck	nose	right cheek	right eye	right jaw	ensemble
ineye botlip	sensitivity	37.75 ± 29.10	27.25 ± 33.93	24.75 ± 25.96	27.00 ± 28.63	37.00 ± 28.72	37.75 ± 31.44	32.00 ± 33.93	43.50 ± 30.32	28.25 ± 27.58	34.50 ± 29.43	37.75 ± 30.47	29.00 ± 27.71
	specificity	72.25 ± 25.95	76.75 ± 25.79	81.00 ± 24.22	79.00 ± 24.9	73.50 ± 23.75	72.25 ± 28.31	77.25 ± 25.79	73.00 ± 24.96	78.50 ± 24.53	77.50 ± 25.88	72.25 ± 23.56	81.25 ± 22.05
	AUC	–	–	–	–	–	–	–	–	–	–	–	0.5770 ± 0.1174
	MCC	0.1120 ± 0.2495	0.0452 ± 0.2612	0.0798 ± 0.2033	0.0767 ± 0.2346	0.1132 ± 0.2187	0.1120 ± 0.2126	0.1038 ± 0.2612	0.1650 ± 0.2428	0.0767 ± 0.2239	0.1315 ± 0.2166	0.1120 ± 0.2764	0.0146 ± 0.2060
	F1-score	46.03 ± 24.93	36.52 ± 27.36	34.86 ± 23.4	37.17 ± 24.66	45.16 ± 26.54	46.03 ± 25.7	41.03 ± 27.36	50.77 ± 26.35	37.17 ± 24.25	43.70 ± 25.25	46.03 ± 26.71	36.97 ± 24.81
outeye nose	sensitivity	30.00 ± 32.04	39.00 ± 31.4	36.00 ± 29.3	29.75 ± 32.43	36.00 ± 32.99	37.50 ± 33.02	39.00 ± 31.4	37.75 ± 32.19	48.00 ± 32.96	32.50 ± 34.16	33.00 ± 34.42	33.25 ± 34.03
	specificity	71.50 ± 27.44	64.50 ± 30.55	69.25 ± 26.63	72.25 ± 31.67	68.25 ± 29.58	66.75 ± 32.5	64.50 ± 30.55	72.75 ± 27.63	63.25 ± 30.77	72.50 ± 28.79	67.25 ± 31.93	71.50 ± 30.20
	AUC	–	–	–	–	–	–	–	–	–	–	–	0.5468 ± 0.1175
	MCC	0.0144 ± 0.2578	0.0408 ± 0.2465	0.0560 ± 0.2510	0.0290 ± 0.2525	0.0418 ± 0.2547	0.0499 ± 0.2262	0.0408 ± 0.2465	0.1120 ± 0.2707	0.1066 ± 0.2804	0.0573 ± 0.2575	0.0000 ± 0.2360	0.0427 ± 0.2502
	F1-score	38.02 ± 27.10	45.11 ± 24.71	42.86 ± 24.78	38.33 ± 26.52	42.52 ± 27.36	44.27 ± 26.35	45.11 ± 24.71	46.03 ± 27.17	51.43 ± 25.74	40.98 ± 27.34	39.68 ± 25.74	40.65 ± 26.91
no align	sensitivity	28.25 ± 32.52	35.75 ± 30.41	30.50 ± 26.39	21.00 ± 27.02	29.75 ± 29.33	31.75 ± 30.33	35.75 ± 30.41	32.50 ± 28.23	39.25 ± 29.45	27.50 ± 31.84	36.25 ± 25.91	27.50 ± 27.55
	specificity	80.75 ± 21.90	74.00 ± 27.30	74.00 ± 24.72	86.00 ± 21.82	76.00 ± 26.23	75.75 ± 26.54	74.00 ± 27.30	75.25 ± 25.69	74.75 ± 24.81	81.50 ± 23.18	74.00 ± 23.94	82.00 ± 20.39
	AUC	–	–	–	–	–	–	–	–	–	–	–	0.5847 ± 0.1438
	MCC	0.0928 ± 0.3031	0.0996 ± 0.2433	0.0438 ± 0.2372	0.0861 ± 0.2335	0.0741 ± 0.3072	0.0883 ± 0.2845	0.0996 ± 0.2433	0.0870 ± 0.2930	0.1548 ± 0.2511	0.1094 ± 0.2879	0.1132 ± 0.2384	0.1094 ± 0.2354
	F1-score	37.50 ± 29.35	43.90 ± 25.73	38.66 ± 23.98	31.07 ± 24.14	39.32 ± 25.91	40.68 ± 26.94	43.90 ± 25.73	41.67 ± 25.52	48.00 ± 25.90	37.84 ± 28.40	45.16 ± 23.38	37.84 ± 25.30

instead apply the affine transformations to individual face regions rather than the whole face itself. This way, anatomical face regions become less varied in their scale. Third, we did not examine which patients were being misclassified across individual face region models. Finally, we selected hyperparameters based on a previous study [14] rather than hyperparameter search. By chance, this could have accidently biased the results towards the proposed model over the conventional models. However, we believe that this chance is relatively small, as it would be unlikely that the same set of hyperparameters leads to optimal performance in the multiple models of each proposed ensemble.

There are two ways in which we may improve upon the features utilized by our proposed method. First, we would like to examine face ratios. Unlike individual landmarks, face ratios would be invariant to scale and may be valuable features in predicting difficult to intubation patients. Such a system would involve detecting pairs of landmarks and computing the ratio of their distance to the distance of another pair of landmarks. Second, and most importantly, we fully intend on utilizing and profile view of patient faces. This aspect is key for features undetectable from front views, mainly having to do with the jaw and neck, and is related to TMD. Similarly, we will analyze front images of patients with their mouths open, which is related to the MP score. Third, we intend to perform analysis on the relationship between features learned by the proposed model and clinical features. Though such an analysis may reveal correlations between deep learning and clinical features, we do not believe that such clinical correlates would positively contribute to preoperative airway assessments, as several multivariable clinical risk formulae based on demographic factors and bedside airway test results [29,30] have failed to perform in a large randomized clinical trial [10]. However, we may still benefit fusing clinical and deep learning features. With these additional changes, we expect that our proposed model's performance can be further improved [31].

We have developed a preliminary method to identify difficult intubation from front face images using an innovative ensemble of CNN-based feature extractors in tandem with attention-based MIL. Our method both exceeds the sensitivity and specificities of conventional bedside tests as well as common deep learning methods. We also demonstrated the importance of utilizing features specific to the face rather than generic features. Through further experimentation, this research will identify facial features that accurately predict difficult intubation. In the future, we will develop a more robust FRFE by augmenting the CASIA-Webface dataset, resolve issues with patient image scale, integrate features related to facial landmark distance ratios, and utilize profile face images in tandem with front views. Successful implementation will result in a model to identify difficult intubation.

Table 4
Results of the FRFE majority voting ten-fold cross-validation using FRFE features. 'ineye botlip' refers to alignment performed using the inner eye corners and bottom lip; 'outeye nose' refers to alignment performed using the outer eye corners and nose; and 'no align' refers to no alignment of the face datasets. Mean and standard deviation are reported. Corresponding confusion matrices can be found in supplementary material.

		bridge	chin	left cheek	left eye	left jaw	mouth	neck	nose	right cheek	right eye	right jaw	ensemble
ineye botlip	sensitivity	67.11 ± 17.81	61.84 ± 17.76	52.63 ± 12.10	56.58 ± 17.20	61.84 ± 26.89	57.89 ± 18.31	55.26 ± 17.07	47.37 ± 18.90	56.58 ± 16.03	63.16 ± 16.37	60.53 ± 18.07	69.74 ± 4.85
	specificity	60.53 ± 18.26	47.37 ± 15.05	56.58 ± 19.05	55.26 ± 23.89	51.32 ± 14.59	50.00 ± 14.07	56.58 ± 23.77	47.37 ± 16.65	56.58 ± 16.73	48.68 ± 21.72	56.58 ± 18.41	64.47 ± 1.84
	AUC	–	–	–	–	–	–	–	–	–	–	–	0.6465 ± 0.0579
	MCC	0.2769 ± 0.2523	0.0931 ± 0.2284	0.0922 ± 0.2618	0.1184 ± 0.3118	0.1323 ± 0.3051	0.0792 ± 0.1833	0.1184 ± 0.2870	−0.0526 ± 0.3029	0.1316 ± 0.2784	0.1197 ± 0.3138	0.1712 ± 0.2569	0.3426 ± 0.0495
	F1-score	64.97 ± 14.07	57.67 ± 15.51	53.69 ± 10.93	56.21 ± 14.58	58.75 ± 21.34	55.70 ± 14.57	55.63 ± 14.29	47.37 ± 16.51	56.58 ± 13.87	58.90 ± 16.14	59.35 ± 13.11	67.95 ± 4.23
outeye nose	sensitivity	52.86 ± 21.35	52.86 ± 22.39	54.29 ± 19.11	57.14 ± 9.04	51.43 ± 19.98	52.86 ± 19.58	60.00 ± 13.13	45.71 ± 22.54	57.14 ± 14.21	54.29 ± 17.10	50.00 ± 22.34	64.29 ± 4.52
	specificity	32.86 ± 20.26	52.86 ± 25.24	61.43 ± 22.13	60.00 ± 13.47	54.29 ± 15.36	55.71 ± 19.11	45.71 ± 28.41	48.57 ± 16.22	58.57 ± 20.20	55.71 ± 18.81	57.14 ± 21.56	58.57 ± 13.55
	AUC	–	–	–	–	–	–	–	–	–	–	–	0.5702 ± 0.0452
	MCC	−0.1476 ± 0.1931	0.0526 ± 0.2931	0.1584 ± 0.3117	0.1712 ± 0.1759	0.0526 ± 0.2740	0.0790 ± 0.2673	0.0665 ± 0.3500	−0.0526 ± 0.2548	0.1579 ± 0.1988	0.0921 ± 0.2186	0.0659 ± 0.3770	0.2372 ± 0.0805
	F1-score	47.90 ± 17.10	52.63 ± 19.24	56.16 ± 15.16	57.72 ± 8.16	52.00 ± 15.88	53.33 ± 13.62	56.44 ± 13.82	46.67 ± 16.77	57.33 ± 9.22	54.30 ± 11.88	51.70 ± 19.29	62.82 ± 1.51
no align	sensitivity	59.21 ± 13.52	60.53 ± 23.80	64.47 ± 13.32	57.89 ± 13.55	56.58 ± 22.59	64.47 ± 22.43	68.42 ± 16.48	56.58 ± 14.51	57.89 ± 17.51	52.63 ± 14.62	63.16 ± 15.82	68.42 ± 9.59
	specificity	50.00 ± 20.16	52.63 ± 19.66	57.89 ± 17.33	47.37 ± 18.83	44.74 ± 13.84	68.42 ± 11.81	61.84 ± 14.97	47.37 ± 18.33	55.26 ± 19.09	50.00 ± 17.27	59.21 ± 19.69	61.84 ± 8.13
	AUC	–	–	–	–	–	–	–	–	–	–	–	0.6331 ± 0.0272
	MCC	0.0925 ± 0.2393	0.1320 ± 0.1880	0.2242 ± 0.1399	0.0529 ± 0.2218	0.0133 ± 0.2449	0.3292 ± 0.2312	0.3033 ± 0.2029	0.0396 ± 0.1851	0.1316 ± 0.1872	0.0263 ± 0.2103	0.2239 ± 0.1654	0.3033 ± 0.0534
	F1-score	56.60 ± 10.60	58.23 ± 16.27	62.42 ± 8.30	55.00 ± 11.32	53.42 ± 17.07	65.77 ± 15.48	66.24 ± 12.04	54.09 ± 10.03	57.14 ± 10.99	51.95 ± 11.65	61.94 ± 9.48	66.24 ± 5.19

Table 5
Results of the MIL majority voting LOO cross-validation. 'ineye botlip' refers to alignment performed using the inner eye corners and bottom lip; 'outeye nose' refers to alignment performed using the outer eye corners and nose; and 'no align' refers to no alignment of the CASIA-Webface dataset. Mean is reported. Corresponding confusion matrices can be found in supplementary material.

	no align mouth	no align left eye	no align left jaw	no align right cheek	ineye botlip neck	ineye botlip chin	ineye botlip left jaw	ineye botlip right jaw	outeye nose chin	ensemble
sensitivity	55.26	55.26	57.89	53.95	51.32	53.95	59.21	59.21	58.44	73.68
specificity	55.26	51.32	60.53	60.53	59.21	55.26	61.84	57.89	60.00	68.42
AUC	–	–	–	–	–	–	–	–	–	0.7105
MCC	0.1053	0.0658	0.1843	0.1451	0.1056	0.0921	0.2106	0.1711	0.1843	0.4216
F1-score	55.26	54.19	58.67	55.78	53.42	54.30	60.00	58.82	58.67	71.79

Funding

This study was partly funded by the Anesthesia Patient Safety Foundation Award 2020 ("Development of machine learning algorithms to predict difficult airway management"), a pilot award provided by Center for Biomedical Informatics at Wake Forest School of Medicine, and NIH R21-EB029493.

Declaration of competing interest

The authors declare no conflicts of interest.

Acknowledgements

We would like to thank CRTs Jacob G. Fowler (B.S.), Easton S. Howard (B.S.), Lauren E. Sands (B.S.), Madeline R. Fram (B.A.), Anthony A. Wachnik (B.S.), Samuel G. Robinson (B.S.), Jessica E. Fanelli (B.S.),

and Nia S. Sweatt (B.S.) for their dedication in acquiring patient images and populating our patient database.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compbiomed.2021.104737.

References

[1] M.E. Detsky, et al., Will this patient be difficult to intubate?: the rational clinical examination systematic review, Jama 321 (5) (2019) 493–503.
[2] C.W. Connor, S. Segal, Accurate classification of difficult intubation by computerized facial analysis, Anesth. Analg. 112 (1) (2011) 84–93.
[3] C.W. Connor, S. Segal, The importance of subjective facial appearance on the ability of anesthesiologists to predict difficult intubation, Anesth. Analg. 118 (2) (2014) 419.
[4] C. Connor, et al., Bedside recruiting and processing OF data ON facial appearance and the ease or difficulty OF intubation, in: ANESTHESIA AND ANALGESIA, 116, LIPPINCOTT WILLIAMS & WILKINS 530 WALNUT ST, PHILADELPHIA, PA 19106-3621 USA, 2013, p. 314, 314.
[5] M.B. Rosenberg, J.C. Phero, Airway assessment for office sedation/anesthesia, Anesth. Prog. 62 (2) (2015) 74–80.
[6] G. Samsoon, J. Young, Difficult tracheal intubation: a retrospective study, Anaesthesia 42 (5) (1987) 487–490.
[7] C. Frerk, Predicting difficult intubation, Anaesthesia 46 (12) (1991) 1005–1008.
[8] T. Shiga, Z.i. Wajima, T. Inoue, A. Sakamoto, Predicting difficult intubation in apparently normal PatientsA meta-analysis of bedside screening test performance, Anesthesiology: The Journal of the American Society of Anesthesiologists 103 (2) (2005) 429–437.
[9] S. Yentis, Predicting difficult intubation–worthwhile exercise or pointless ritual? Anaesthesia 57 (2) (2002) 105.
[10] A.K. Nørskov, et al., Effects of using the simplified airway risk index vs usual airway assessment on unanticipated difficult tracheal intubation-a cluster randomized trial with 64,273 participants, Br. J. Addiction: Br. J. Anaesth. 116 (5) (2016) 680–689.
[11] C.W. Connor, S. Segal, Systems and Methods for Predicting Potentially Difficult Intubation of a Subject, 2013.
[12] G.L. Cuendet, et al., Facial image analysis for fully automatic prediction of difficult endotracheal intubation, IEEE (Inst. Electr. Electron. Eng.) Trans. Biomed. Eng. 63 (2) (2015) 328–339.
[13] I. Goodfellow, Y. Bengio, A. Courville, Y. Bengio, Deep Learning, MIT press Cambridge, 2016.
[14] Y. Gurovich, et al., Identifying facial phenotypes of genetic disorders using deep learning, Nat. Med. 25 (1) (2019) 60–64.
[15] S.J. Pan, Q. Yang, A survey on transfer learning, IEEE Trans. Knowl. Data Eng. 22 (10) (2009) 1345–1359.
[16] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: a large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, Ieee, 2009, pp. 248–255.
[17] M.K.K. Niazi, T.E. Tavolara, V. Arole, D.J. Hartman, L. Pantanowitz, M.N. Gurcan, Identifying tumor in pancreatic neuroendocrine neoplasms from Ki67 images using transfer learning, PloS One 13 (4) (2018), e0195621.
[18] T. E. Tavolara, M. K. K. Niazi, W. Chen, W. Frankel, and M. N. Gurcan, "Colorectal tumor identification by transferring knowledge from pan-cytokeratin to H&E," vol. 10956, p. 1095614: International Society for Optics and Photonics.
[19] D. Yi, Z. Lei, S. Liao, S.Z. Li, Learning Face Representation from Scratch, arXiv preprint arXiv:, 2014.
[20] M. Ilse, J.M. Tomczak, M. Welling, Attention-based deep multiple instance learning, 2018 arXiv preprint arXiv:.04712.
[21] D.E. King, Dlib-ml: a machine learning toolkit, J. Mach. Learn. Res. 10 (2009) 1755–1758.
[22] V. Kazemi, J. Sullivan, One millisecond face alignment with an ensemble of regression trees, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1867–1874.
[23] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition 1, CVPR'05, 2005, pp. 886–893. Ieee.
[24] B. Amos, B. Ludwiczuk, M. Satyanarayanan, Openface: a general-purpose face recognition library with mobile applications, CMU School of Computer Science 6 (2) (2016).
[25] K. He, X. Zhang, S. Ren, J. Sun, Delving deep into rectifiers: surpassing human-level performance on imagenet classification, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1026–1034.
[26] O. Maron, T. Lozano-Pérez, A framework for multiple-instance learning, Adv. Neural Inf. Process. Syst. (1998) 570–576.
[27] C. Szegedy, S. Ioffe, V. Vanhoucke, A. Alemi, Inception-v4, inception-resnet and the impact of residual connections on learning, 2016 arXiv preprint arXiv:.07261.
[28] A. Trajman, R.R. Luiz, McNemar $\chi 2$ test revisited: comparing sensitivity and specificity of diagnostic examinations, Scandinavian journal of clinical and laboratory investigation 68 (1) (2008) 77–80.
[29] M. Naguib, et al., Predictive performance of three multivariate difficult tracheal intubation models: a double-blind, case-controlled study, Anesth. Analg. 102 (3) (2006) 818–824.
[30] J. L'Hermite, E. Nouvellon, P. Cuvillon, P. Fabbro-Peray, O. Langeron, J. Ripart, The Simplified Predictive Intubation Difficulty Score: a new weighted score for difficult airway assessment, European Journal of Anaesthesiology| EJA 26 (12) (2009) 1003–1009.
[31] J.T. Roberts, H.H. Ali, G.D. Shorten, Using the laryngeal indices caliper to predict difficulty of laryngoscopy with a Macintosh# 3 laryngoscope, J. Clin. Anesth. 5 (4) (1993) 302–305.