ANNALS of Clinical and Translational Neurology
Open Access
AMERICAN NEUROLOGICAL ASSOCIATION

RESEARCH ARTICLE

Assessing facial weakness in myasthenia gravis with facial recognition software and deep learning

Annabel M. Ruiter[1], Ziqi Wang[2], Zhao Yin[2], Willemijn C. Naber[1], Jerrel Simons[1], Jurre T. Blom[3], Jan C. van Gemert[2], Jan J. G. M. Verschuuren[1] & Martijn R. Tannemaat[1]

[1]Department of Neurology, Leiden University Medical Center, Leiden, the Netherlands
[2]Vision Lab, Delft University of Technology, Delft, the Netherlands
[3]Medical Illustrator at www.jurreblom.nl, Apeldoorn, the Netherlands

Correspondence
Annabel Ruiter, Department of Neurology, Leiden University Medical Center, Albinusdreef 2, 2333 ZA, Leiden, the Netherlands. Tel: +31-71-5262197; Fax +31-71-5266671; E-mail: a.m.ruiter@lumc.nl

Received: 16 March 2023; Revised: 23 May 2023; Accepted: 24 May 2023

Annals of Clinical and Translational Neurology **2023; 10(8): 1314–1325**

doi: 10.1002/acn3.51823

Abstract

Objective: Myasthenia gravis (MG) is an autoimmune disease leading to fatigable muscle weakness. Extra-ocular and bulbar muscles are most commonly affected. We aimed to investigate whether facial weakness can be quantified automatically and used for diagnosis and disease monitoring. **Methods**: In this cross-sectional study, we analyzed video recordings of 70 MG patients and 69 healthy controls (HC) with two different methods. Facial weakness was first quantified with facial expression recognition software. Subsequently, a deep learning (DL) computer model was trained for the classification of diagnosis and disease severity using multiple cross-validations on videos of 50 patients and 50 controls. Results were validated using unseen videos of 20 MG patients and 19 HC. **Results**: Expression of anger ($p = 0.026$), fear ($p = 0.003$), and happiness ($p < 0.001$) was significantly decreased in MG compared to HC. Specific patterns of decreased facial movement were detectable in each emotion. Results of the DL model for diagnosis were as follows: area under the curve (AUC) of the receiver operator curve 0.75 (95% CI 0.65–0.85), sensitivity 0.76, specificity 0.76, and accuracy 76%. For disease severity: AUC 0.75 (95% CI 0.60–0.90), sensitivity 0.93, specificity 0.63, and accuracy 80%. Results of validation, diagnosis: AUC 0.82 (95% CI: 0.67–0.97), sensitivity 1.0, specificity 0.74, and accuracy 87%. For disease severity: AUC 0.88 (95% CI: 0.67–1.0), sensitivity 1.0, specificity 0.86, and accuracy 94%. **Interpretation**: Patterns of facial weakness can be detected with facial recognition software. Second, this study delivers a 'proof of concept' for a DL model that can distinguish MG from HC and classifies disease severity.

Introduction

Myasthenia gravis (MG) is an autoimmune disease characterized by fluctuating muscle fatigability.[1] The prevalence of MG is low: approximately 1 to 2 per 10,000.[1] Although all striated muscles can be involved, the extra-ocular muscles are most commonly affected.[1,2] Fluctuating asymmetric ptosis and diplopia are the first and predominant symptoms in a majority of patients, followed by weakness of bulbar muscles, resulting in facial weakness, amongst other symptoms.[1,3] The diagnosis of MG is based on typical clinical features in combination with the presence of auto-antibodies against neuromuscular junction proteins, abnormal clinical neurophysiological tests or a positive neostigmine test.[4] Despite its typical semiology and a range of ancillary tests, diagnosing MG can be challenging. Up to 46% of MG patients do not receive the correct diagnosis within the first year of onset.[5]

In addition, the fluctuating day-to-day severity of their muscle weakness is an issue of great concern for all patients with MG. Disease exacerbations, which can involve impaired breathing, leading to life-threatening situations, are almost always preceded by increased bulbar muscle weakness or worsening of primary symptoms.[6] These concerns are compounded by the fact that care for

rare diseases is increasingly clustered in expert centers, leading to prolonged travel times to the hospital. A personal digital disease monitoring tool could improve outcomes by tracking symptoms and quality of life.[7] Such a tool could potentially be used for early detection or even prevention of exacerbations by allowing patients to make timely adjustments to their immune suppressant medication.[8]

Artificial intelligence (AI) applications are an area of intense medical research, particularly to improve diagnosis and for home monitoring of disease severity.[9] However, to our knowledge AI applications are relatively scarce in the field of neuromuscular disorders so far and appear to be restricted to muscle imaging and gene profiling.[10–14] Assessment of facial weakness has thus far only been carried out in facial palsies and almost exclusively on photographs.[15–20] As the facial features of MG patients are distinct from healthy subjects,[21] we hypothesized that a short video recording of MG patients contains information that can be used both for diagnostic and monitoring purposes.

We aimed to (1) quantify and map patterns of facial weakness with the use of facial expression recognition software and (2) to develop a deep learning (DL) computer model for diagnosis and disease monitoring.

Materials and Methods

This study was conducted between May 2019 and September 2020 at the Department of Neurology of the Leiden University Medical Center (LUMC), Leiden, the Netherlands, and the Technical University Delft, the Netherlands.

Standard protocol approvals, registrations, and patient consents

The ethics committee of the LUMC approved the study protocol. Written consent was obtained from all participants according to the declaration of Helsinki. A signed patient consent form has been obtained for the publication of recognizable photographs.

Participants

Seventy MG patients and 69 age- and gender-matched healthy controls (HC) were recruited from the Neurology outpatient clinic of the LUMC. Inclusion criteria were: age $\geq$ 18 years. Diagnosis of MG was based on clinical signs or symptoms supportive of MG and at least one of the following: a positive serologic test for AChR or MuSK antibodies and/or a diagnostic electrophysiological investigation supportive of the diagnosis myasthenia gravis and/or a positive neostigmine test. Participants were excluded if they were unable to give written informed consent or read Dutch or English video instructions. Second, the presence of other medical conditions affecting the facial muscles was an exclusion criteria: e.g., active Graves' disease or unilateral facial paralysis. The use of corticosteroids was an exclusion criteria in the healthy control group.

All participants were included for quantification and mapping of facial weakness with the use of facial expression recognition software.

For the development of a DL model, the first 50 MG patients and 50 HC were selected for training of the model. For external validation, the remaining 20 MG patients and 19 HC were used.

Procedures

Videos were recorded with a 4K-camera in a standardized setting with the assistance of the researcher, present during the entire recording. The subject was seated in front of a green screen with sufficient lighting. A computer screen displayed a slideshow of a model portraying 24 different facial expressions, which participants were asked to mimic with maximal intensity for the duration of each slide. The slideshow was accompanied by written Dutch instructions; verbal instructions were given when necessary. The slideshow contained the following expressions: neutral, eyes closed, right/left eye closed, eyes closed firmly, anger, fear, happiness, surprise, disgust, sadness, showing teeth, raising eyebrows, mouth open, whistling, tongue protrusion (8 sec each) and sustained gaze in eight directions (30 sec for left and right gaze, 4 sec for the other directions). Disease severity was measured using the quantitative MG score (QMG). All QMGs were performed approximately 1 h before the recording by an experienced nurse who was trained and certified to perform QMG assessments. A QMG score of 0–9 was classified as mild and a QMG score >9 as moderate–severe disease.[22] The first five items of the QMG score were used as a degree of facial weakness (score 0–15). The current MG Foundation of America (MGFA) A (limb predominant) or B (bulbar predominant) classification was retrieved from patients' medical files. The cumulative prednisone dosage over the 6 months prior to participation was calculated because of the potential confounding effect of facial changes due to corticosteroids.

Quantification and pattern mapping of facial weakness

Quantification and mapping of facial weakness were assessed using FaceReader facial expression recognition

software version 8 (Noldus Information Technology BV[23]) on the original video (resolution 3840×2160 pixels). This software classifies and quantifies the perceived expression of six emotions: anger, fear, happiness, surprise, disgust, and sadness, by generating a numeric value between 0 and 1, every 80 ms throughout the video. Higher values correspond with more pronounced expression. The software also quantifies 20 individual action units (AUs, Table S1) in a similar manner. AUs are part of the Facial Action Coding System (FACS),[24] developed for standardized assessment of facial behavior. Individual AUs correspond with a specific facial movement, which is related to the activation of a certain muscle (E.g. AU1, movement: inner brow raiser, muscle: m. frontalis pars medialis. Table S1). The AU represents both the left and right sides of the face.

For this study, we compared mean scores between MG patients and HC of all six emotions, when prompted to mimic the corresponding emotion. Second, we compared the median scores of all 20 individual AUs between the two groups during the expression of these six emotions. Both analyses were subsequently performed for disease severity.

Development of deep learning model

We used a three-dimensional (3D) convolutional neural network (CNN). The convolutional kernel size in a two-dimensional (2D) CNN is $N \times N$, corresponding to the width and height ($W \times H$) of the image filter, where N determines the perception field in 2D. These 2D kernels can only capture spatial information in an image, which can be sufficient for many large-scale action recognition tasks.[25,26] As muscle weakness temporally fluctuates in MG, we added temporal information as a third dimension, yielding a 3D CNN. We adopted an commonly used inflated 3D network (I3D)[27,28] for capturing information in the temporal dimension T. The original settings were unaltered and available from the original sources.[27,28] Different from a 2D CNN, the convolutional kernel in an I3D network is inflated to be $N \times N \times N$, which matches three dimensions: time, width, and height ($T \times W \times H$). We evaluated three 2D CNN backbones into I3D architectures, namely ResNet-18, ResNet34, and ResNet-50.[29] The numbers indicate the depth of the ResNet architecture. Learned features of videos were flattened by a fully connected layer and a *Softmax* activation function was used to obtain the posterior probabilities $P(\hat{y}|x)$ of an input x. Posteriors are the output probabilities for each class of the neural network. The Cross-entropy loss measured the quality of prediction $\hat{y}$ by comparing it to the ground truth label y. The model was optimized by backpropagating the gradient of the Softmax loss function: $L_{\hat{y}\ y} = -\sum_c^n yc \log(P(\hat{y}_c|x))$, where c indicates the class label and n the number of classes. y_c is the ground truth binary indicator of class c for input x.

To prevent overfitting, the model was pretrained using the Kinetics dataset.[26] Videos were cut into clips containing one specific expression manually by human visual assessment. Transition times between performed tasks were deleted. Duration of individual clips varied from 2 to 33 sec, with a mean of 6 sec. Resolution of clips was 256×256 pixels and no cropping was applied. Clips were processed per 16 batches (i.e., the gradients were aggregated and updated over 16 batches) and each batch contained 64 frames (frame rate video 32/sec), as the number of frames that can be processed in a batch is limited by the computational memory. Each batch contained frames from a single clip. Patient-level prediction was determined by the majority vote of all video clips per patient. Performance of the network is sensitive to hyperparameters[30–32]; we adopted the stochastic gradient descent with momentum as optimizer with the learning rate of 0.01 and weight decay of 0.0001. We set the momentum parameter at 0.01.

For training, a threefold cross-validation with a 2:1 split was applied for the classification of diagnosis, with MG patients as the true class. For disease severity, a fourfold cross-validation with a 3:1 split was applied, with patients with QMG 0–9 as the true class. After completing the training phase, the model was applied to a completely new and unseen dataset of videos without any further modifications.

Comparison of the deep learning model and neurologists

Four neurologists specialized in neuromuscular disorders each rated 50 videos for diagnosis using a visual analog scale, resulting in a probability score between zero and one. They subsequently each rated 20 videos for disease severity (mild or moderate–severe disease) in a similar way. Videos were selected semi-randomized by the researcher: at least 20 healthy controls were included in the dataset for diagnosis. If a neurologist was familiar with a subject, this specific video was excluded. Datasets, therefore, differed between neurologists, as different patients were excluded by each neurologist. The inter-rater reliability (IRR) scores were calculated for videos rated by two or all four neurologists. Mean scores were calculated if subjects were rated by more than one neurologist. The results of all four neurologists were pooled for comparison with the results of the deep learning model.

Statistical analysis

Quantification of six emotions and 20 different AUs every 80 ms with FaceReader resulted in 100 values between 0 and 1 for each individual AU and the emotion during the 8 sec that this expression was maintained. This data was transcribed to an expression-based time-weighted table in MATLAB (2020, MathWorks). To correct for potential outliers, 95th percentile scores of emotions were calculated during the performance of the corresponding emotion, e.g., during expression of happiness, we calculated the 95th percentile score of the FaceReader-derived score for "happiness." Second, the 95th percentile score of each individual AU was calculated during the performance of each emotion. Per video, this resulted in one value for each emotion and six values for each of the 20 individual AU (one for each of the six emotions).

Statistical analyses were performed using IBM SPSS Statistics version 25.0. Group data are described by mean and standard deviation (±SD) for normally distributed data or median and interquartile range [IQR] for all other data. Unpaired T-tests or Mann–Whitney U tests were used for comparison between two groups, one-way ANOVA for >2 groups. Significance was accepted at $p < 0.05$. Performance was determined by creating a receiver-operating characteristics (ROC) curve and determining the area under the curve (AUC), sensitivity, specificity, and accuracy. A logistic regression analysis with all FaceReader variables that showed significant differences in the univariate analysis was performed to compare the overall performance of FaceReader with the deep learning model. The design of this study is exploratory, therefore we did not apply post-hoc correction for multiple testing.

Results

Baseline characteristics

Baseline characteristics are presented in Table 1. Baseline characteristics of participants did not differ between

Table 1. Baseline characteristics, grouped according to the groups used for the deep learning model.

	Patients training	Controls training	Patients validation	Controls validation	Sig.
Participants, *n* (% male)	50 (38%)	50 (38%)	20 (35%)	19 (42%)	–
Age, mean ± SD	54.7 ± 18.6[b]	49.4 ± 15.1[b]	55.3 ± 14.6[c]	48.1 ± 11.8[c]	0.910[a]; 0.731[a]
Disease duration in years, median [IQR]	5.8 [2.4–18.8]	–	4.5 [1.8–19.2]	–	0.687
Antibodies, %		–		–	–
AChR	72%		90%		
MuSK	14%		5%		
Seronegative	14%		5%		
Total QMG score, mean ± SD	9.0 ± 4.8	–	11.0 ± 6.7	–	0.274
Missing, *n*	1		2		
Facial QMG score, mean ± SD	2.8 ± 2.4		3.4 ± 3.4		0.493
Facial QMG score 0–1 points, %	38%		41%		
Mild MG (QMG 0–9)					
N, %	29 (59%)		10 (56%)		
QMG, mean ± SD	5.6 ± 2.3		6.0 ± 1.9		0.648
MGFA A/B, %	3%/34%		20%/10%		
Moderate–severe MG (QMG >9)					
N, %	20 (41%)		8 (44%)		
QMG, mean ± SD	13.8 ± 3.2		17.5 ± 4.2		**0.019**
MGFA A/B, %	20%/65%		25%/63%		
Cumulative prednisone dosage (mg) past 6 months, median [IQR]		–		–	
Mild	827.5 [0–2151][d]		0 [0–0][e]		**0.003**
Moderate–severe	1471.6 [0–3758][d]		1362.5 [0–3356][e]		0.833

Bold values represent *p*-values <0.05.

MG, myasthenia gravis; MGFA, myasthenia gravis foundation of America (A = limb predominant, B = bulbar predominant) at the time of the video recording; QMG, quantitative myasthenia gravis score.

[a]First value represents between patient sets, second value is between control sets.

[b]No significant difference in age between patients vs controls in training set: $p = 0.117$.

[c]No significant difference in age between patients vs controls in validation set: $p = 0.099$.

[d]No significant difference in prednisone dosage between mild and moderate–severe patients in training set: $p = 0.323$.

[e]No significant difference in prednisone dosage between mild and moderate–severe patients in the validation set: $p = 0.079$.

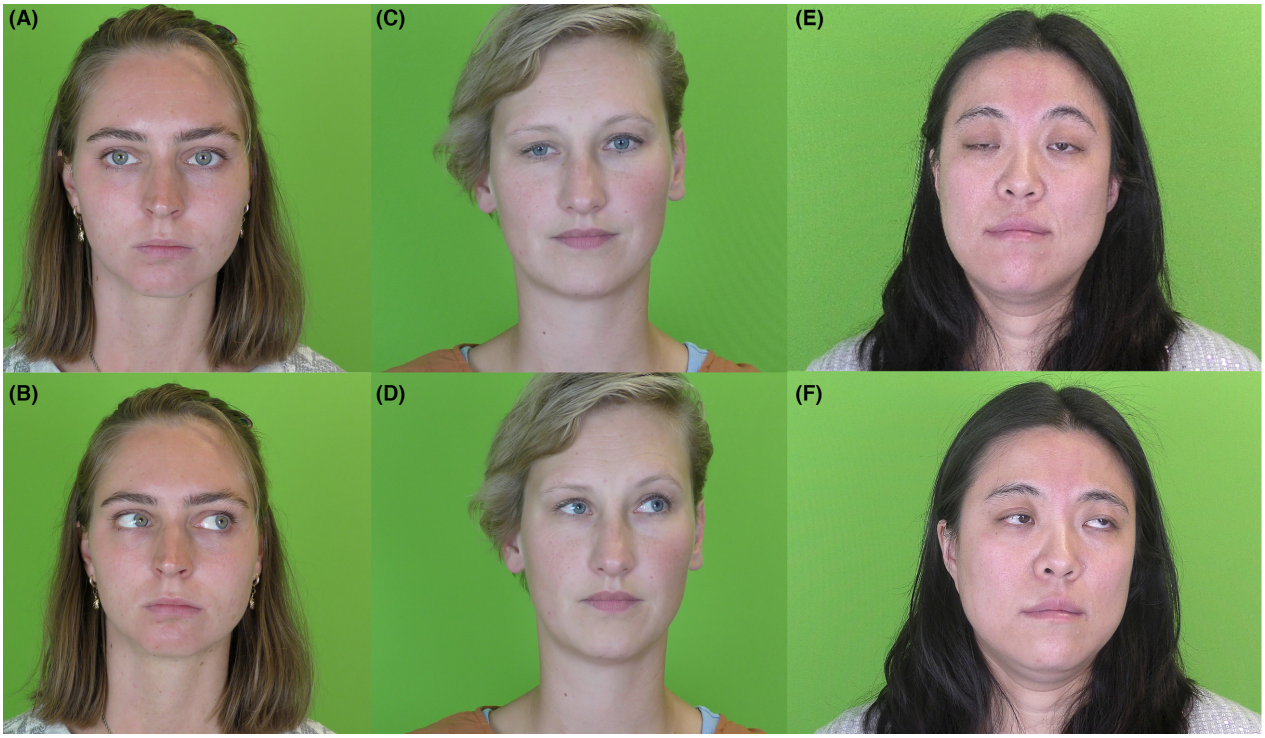

Figure 1. Video screen shots of three different participants. (A and B) Healthy control; (C and D) MG patient with mild disease (total QMG score 9): slight right-sided ptosis; (E and F) MG patient with moderate–severe disease (total QMG score 24): severe bilateral ptosis, compensatory raising of eyebrows, ophthalmoplegia, lower facial weakness.

videos assessed by the deep learning model and neurologists or between individual neurologists (Table S2). Figure 1A–F shows screenshots of two expressions (neutral and left gaze) of three different participants: an HC, an MG patient with mild disease, and one with moderate–severe disease.

Quantification and mapping of facial weakness

The mean FaceReader-derived scores of the emotions anger, fear, and happiness were significantly lower during the expression of these specific emotions in patients with MG, compared to HC (Table 2, Fig. 2).

The AUC of the ROC curve of "anger" was 0.60 (95% CI 0.51–0.70). Sensitivity was 0.66, specificity was 0.54, and accuracy reached 58%. For "fear" the AUC of the ROC curve was 0.64 (95% CI 0.55–0.74). Sensitivity was 0.62, specificity 0.62, and accuracy reached 54%. For "happiness" the AUC was 0.71 (95% CI 0.62–0.79). Sensitivity was 0.75, specificity was 0.65, and accuracy reached 70%. There were no significant differences in mean scores of expressed emotions between patients with mild and moderate–severe disease (Table 2). There was, however, a significant negative correlation between the expression of "disgusted" and the QMG score (−0.28, $p = 0.045$).

Figure 3 summarizes all AUs that differed significantly between MG and HC during the expression of the six emotions. Raw scores are available in Table S3.

The following differences were observed between MG patients and HC. During the expression of anger, brows were lowered less (AU4), lips were pressed, and tightened less (AU23, 24) and eyelids were lowered more (AU43). During the expression of fear: brows were lowered less (AU4), eyelids were tightened more (AU7), the upper lip was raised more (AU10) and eyelids were lowered more (AU43). During the expression of happiness: cheeks were raised less (AU6), lip corners were pulled up less (AU12) and lips parted less (AU25). During the expression of surprise: the outer parts of the brows were raised less (AU2), cheeks were raised less (AU6), lip corners were pulled up less (AU12), lips parted less (AU25), and the mouth was stretched more (AU27). During the expression of disgust: brows were lowered less (AU4) and eyelids were lowered more (AU43). Sadness: eyelids were lowered more (AU43). Logistic regression on the emotions and AUs with a significant difference between MG and HC, yielded

Table 2. Quantification of six emotions in myasthenia gravis patients versus healthy controls (upper part table) and mild versus moderate–severe disease (lower part table).

Facial expression	Myasthenia gravis patients ($n = 70$)	Healthy controls ($n = 69$)	p-value
Anger, mean ± SD	0.33 ± 0.26	0.45 ± 0.32	**0.026**
Fear, mean ± SD	0.12 ± 0.21	0.24 ± 0.27	**0.003**
Happiness, mean ± SD	0.59 ± 0.35	0.80 ± 0.27	**<0.001**
Surprise, mean ± SD	0.44 ± 0.34	0.40 ± 0.34	0.422
Disgust, mean ± SD	0.42 ± 0.31	0.45 ± 0.35	0.604
Sadness, mean ± SD	0.42 ± 0.34	0.49 ± 0.34	0.263

Facial expression	Mild disease ($n = 39$)	Moderate severe disease ($n = 28$)	p-value
Anger, mean ± SD	0.32 ± 0.24	0.33 ± 0.28	0.906
Fear, mean ± SD	0.15 ± 0.25	0.09 ± 0.14	0.224
Happiness, mean ± SD	0.65 ± 0.32	0.53 ± 0.37	0.160
Surprise, mean ± SD	0.46 ± 0.35	0.43 ± 0.35	0.750
Disgust, mean ± SD	0.49 ± 0.29	0.34 ± 0.32	0.054
Sadness, mean ± SD	0.41 ± 0.33	0.43 ± 0.34	0.837

Bold values represent p-values <0.05.

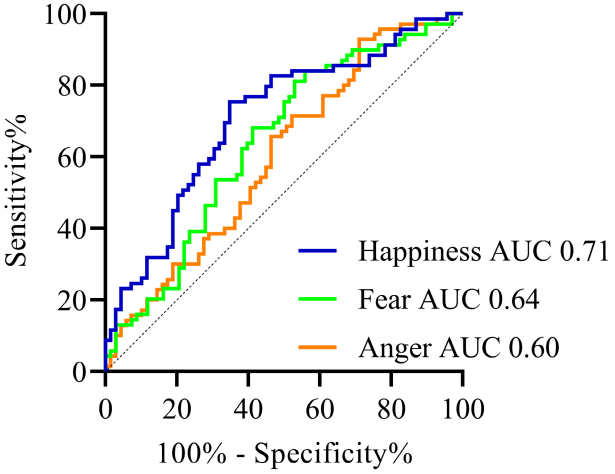

Figure 2. Receiver-operating curves of emotions anger, fear, and happiness in myasthenia gravis compared to healthy controls.

a accuracy of 77%. The extensive output table, including individual odd ratios, is available as Table S4.

When comparing patients with moderate–severe disease to mild disease (Table S3): during expression of happiness: cheeks raised less (AU6) and lip corners pulled up less (AU12). Disgust: lips parted less (AU25). Sadness: cheeks raised less (AU6). A negative correlation with QMG score was present for: anger AU6 (−0.274, $p = 0.025$), happiness AU6 (−0.326, $p = 0.007$), surprise AU6 (−0.293, $p = 0.017$), sadness AU6 (−0.306, $p = 0.012$). Logistic regression on the AUs with a significant difference between the two severity groups, yielded an accuracy of 77%. The extensive output table, including individual odd ratios, is available as Table S5.

Deep learning model

Highest performance was achieved using a 3D ResNet-50 for diagnosis and a 3D ResNet-34 for disease severity. Results of the multiple cross-validation on the training set were as follows: for diagnosis, the AUC was 0.75 (95% CI 0.65–0.85). Sensitivity and specificity were both 0.76, accuracy reached 76%. For disease severity, the AUC was 0.75 (95% CI 0.60–0.90). Sensitivity 0.93, specificity 0.63, and accuracy reached 80%. Of the incorrect classifications, 78% were incorrectly classified as mild disease severity. According to their MGFA classification: 1 (14%) was class I (pure ocular), 29% class A and 57% was class B ($p = 0.828$). The remaining 22% was incorrectly classified as moderate–severe disease severity. All were MGFA class 0 (remission) according to their medical file. Overall, 44% of incorrect classifications occurred in MGFA class B, 33% in class 0/1, and 22% in class B ($p = 0.450$).

Application of the trained model to an unseen validation set of MG patients and HC yielded the following results (Fig. 4): for diagnosis, AUC was 0.82 (95% CI: 0.67–0.97). Sensitivity 1.0, specificity 0.74, and accuracy reached 87%. AUC for disease severity was 0.88 (95% CI: 0.67–1.0); sensitivity was 1.0, specificity 0.86, and accuracy reached 94%. Only one incorrect classification occurred in the moderate–severe disease severity group, which was an MGFA class A patient.

Neuromuscular neurologists

Mean probability scores were calculated from the results of all four neurologists combined. For diagnosis, AUC was 0.71 (95% CI: 0.61–0.82) (Fig. 4). Sensitivity and

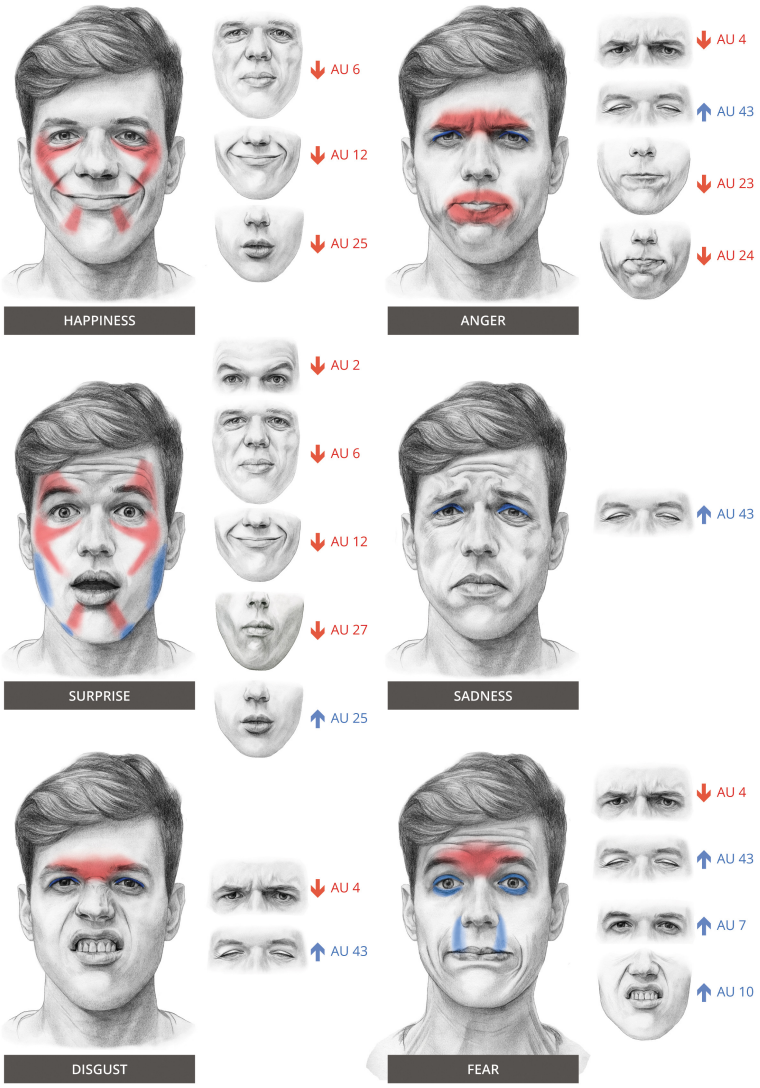

Figure 3. Patterns of weakness in myasthenia gravis during the expression of six emotions. In red: decrease in AU compared to HC, and in blue: an increase in AU compared to HC. Smaller images on the right of each emotion show the individual AU with the corresponding movement.

specificity were 0.69 and 0.67, respectively; accuracy reached 68%. AUC for disease severity was 0.63 (95% CI: 0.47–0.78). Sensitivity and specificity were 0.68 and 0.57, respectively, and accuracy reached 63%. Of the incorrect classifications, 71% were incorrectly classified as mild disease severity. According to their MGFA classification: 35% was class 0 or I, 20% was class A, and 45% was class B ($p = 0.368$). The remaining 29% was incorrectly classified as moderate–severe disease severity. According to their MGFA classification: 50% was class 0 or I, 13% was class A, and 38% was class B ($p = 0.461$). Overall, 43% of incorrect classification occurred in MGFA class B, 39% in class 0/I, and 18% in class B ($p = 0.125$).

Performances of individual neurologists are displayed in Figure 5 and Table S6. For the classification of diagnosis, only 18 videos were assessed by all four neurologists. The IRR for this subset was 0.214. For disease severity, no videos were rated by all four neurologists. IRR for videos rated by two neurologists are displayed in Table 3.

Discussion

To our knowledge, this is the first study using advanced computer analytics to quantify facial weakness in MG. We show that in a large group of chronic MG patients with relatively low QMG scores, the overall weakness of

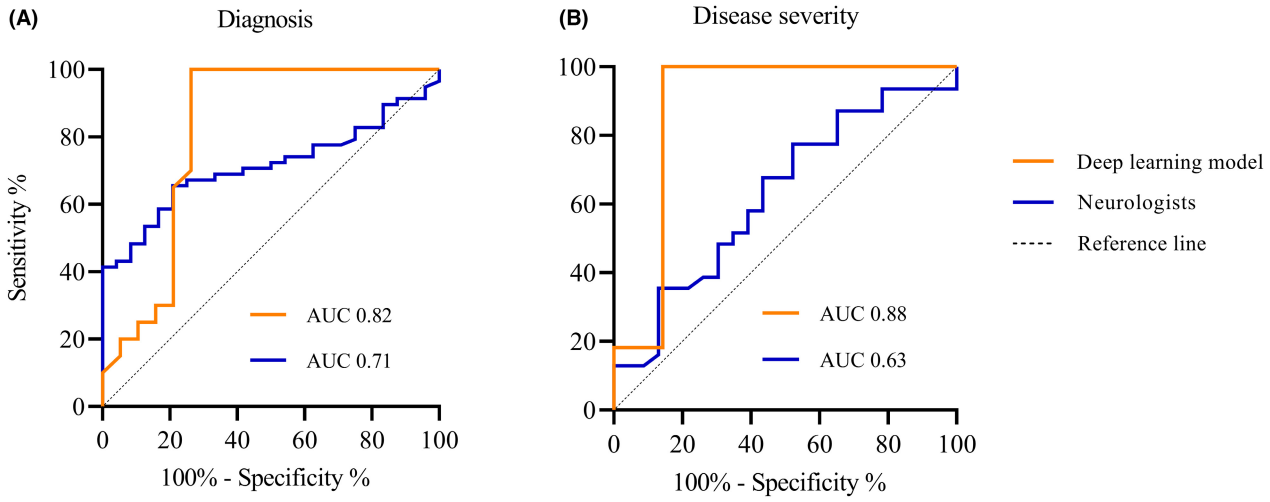

Figure 4. ROC curves deep learning model and neurologists. (A) results for diagnosis, (B) results for disease severity.

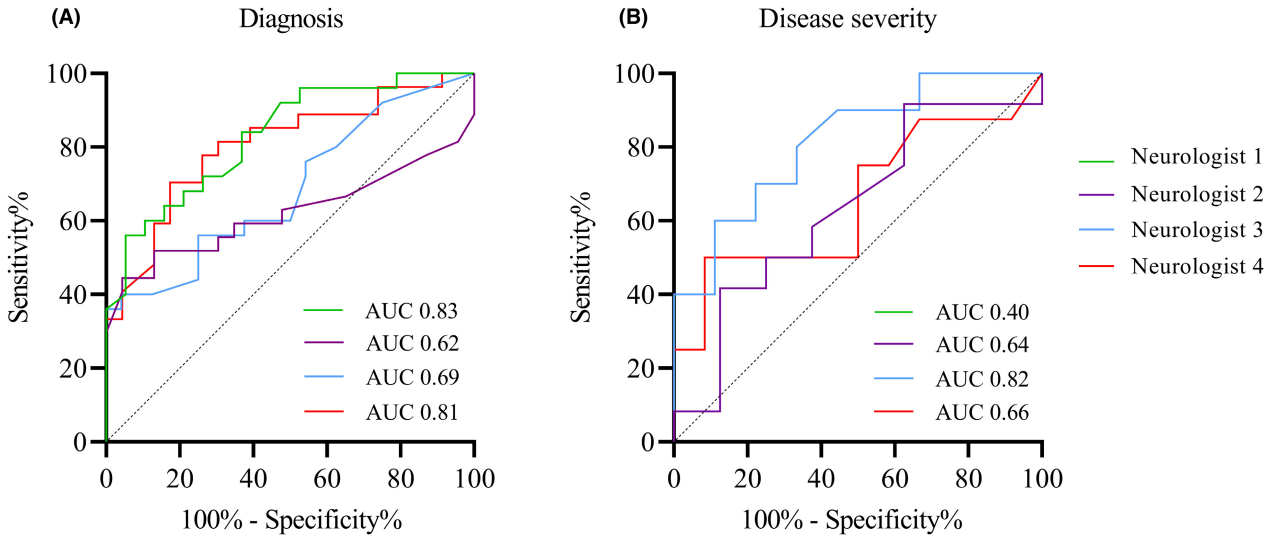

Figure 5. ROC curves individual neurologists. (A) results for diagnosis, (B) results for disease severity. Neurologist 1 is not plotted in B because of AUC <0.5.

Table 3. Interrater-reliability scores for videos rated by two neurologists.

	Neurologist 1	Neurologist 2	Neurologist 3	Neurologist 4
Neurologist 1	NA	0.418 ($n = 29$)	0.138 ($n = 21$)	0.528 ($n = 29$)
Neurologist 2	NA ($n = 1$)	NA	0.513 ($n = 29$)	0.592 ($n = 50$)
Neurologist 3	NA ($n = 0$)	−0.244 ($n = 4$)	NA	0.584 ($n = 29$)
Neurologist 4	NA ($n = 1$)	0.316 ($n = 20$)	0.239 ($n = 4$)	NA

Abbreviation: NA, not applicable.
The upper right part is for classification of diagnosis, lower left part is for classification of disease severity. N = the number of videos rated by both neurologists.

facial muscles is a common occurrence. This is in accordance with previous research, which showed that more than 60% of MG patients reported facial weakness, dysarthria, and/or weakness with chewing or swallowing.[3] This weakness can be quantified with the use of facial expression recognition software[23] and could potentially serve as a diagnostic tool and a biomarker for disease severity. In addition, a DL model trained on video data of facial expressions was capable of recognizing patients with MG and classifying disease severity with high accuracy.

Quantification of facial weakness

MG patients had a reduced expression of the emotions of anger, fear, and happiness, most likely due to weakness of the facial muscles. Facial expression is similarly affected in facial palsies and facioscapulohumeral dystrophy and affects social interaction.[15,33,34] The effect of impaired facial expression on non-verbal communication in MG has not been previously studied. However, a negative impact on daily interaction and communication is to be expected. In five out of six emotions, activity was decreased in one or more AUs necessary to maintain that specific expression, e.g.: pulling of the lip corners was less pronounced during the expression of happiness (AU12). Typical MG facial weakness was present in two-thirds of emotions: an increase in AU43 (eye closure), which was likely affected by ptosis. Unfortunately, FaceReader was not able to provide information on gaze deviations, which typically occur during persistent gaze.[35]

FaceReader showed few differences between mild and moderate–severe disease, but AU6, corresponding to action of the cheek raiser or m. orbicularis oculi was significantly lower during two emotions in moderate–severe compared to mild patients and negatively correlated with the QMG score in four emotions. Weakness of the m. orbicularis oculi is common in MG[2] and AU6 could therefore be a potential biomarker for disease severity. However, we did not correct for multiple testing in this exploratory study and this would require further validation in a longitudinal study.

Comparison with previous studies on facial weakness

Previous publications on facial imaging in neuromuscular disorders have focused on facial palsies[15–18,36] and commonly used photographs. Only one study used facial expression analysis software on video data for the quantification of basic facial expressions.[15] Facial imaging has also been used for the diagnosis of genetic disorders.[37–39] In these studies, photographs were used to identify specific disorders based on typical facial dysmorphias. To our knowledge, no previous published study has made use of the temporal dimension in videos for the development of a DL model to assess facial movement and weakness in neuromuscular disorders. The main benefit of using videos over photos is the ability to capture typical effort-dependent weakness of facial muscles in MG. Ptosis is present in 66% of MG patients. It takes, on average, 28 seconds to develop[35] and it is more likely to become apparent in video recordings than in photographs.

For both diagnosis and classification of disease severity, the DL model performed better than four specialized neuromuscular neurologists, combined or individually. It should be noted that we asked both the model and the neurologists to assess an isolated phenomenon without any clinical context: a video of facial expressions. This differs significantly from clinical practice, in which a detailed history, physical examination, and ancillary tests are essential elements for establishing the diagnosis. This is both a strength and weakness of our study. On one hand, it shows that assessment of a facial video alone could potentially be sufficient to reach a diagnosis or estimate disease severity, although this would likely require an improved model, based on data from different populations and centers and a model trained to distinguish MG patients from disease controls. On the other hand, the inclusion of additional information would likely lead to even better results and would result in a more relevant comparison of the diagnostic performance of the DL model and clinical experts.

Interestingly, the diagnostic accuracy of the neurologists assessing facial videos was lower and showed more variation than expected. This may have been caused by the fact that neurologists were asked to perform an unfamiliar task: to assess an isolated phenomenon on a standardized video with no sound without any clinical context such as the patients history, physical examination, or results from additional investigations. There were no differences in the occurrence of incorrect classifications between patients with a limb or bulbar predominant disease pattern, not for the DL model as well as the clinical experts.

Applying a multivariate logistic regression analysis on all emotions and AUs with a significant difference, yielded similar results as the DL model during training: 77% versus 76% for diagnosis and 77% versus 80% for disease severity. However, validation of the DL model on unseen videos achieved higher accuracies. It would be of interest if the FaceReader software could detect differences in weakness over time.

Potential clinical applications

For diagnostic purposes, the results of the DL model should be considered a 'proof of concept' as we included

a relatively small number of MG patients from one center and no patients with other causes of facial weakness or 'MG-mimics', with and without steroids, such as facial nerve palsy, oculopharyngeal muscular dystrophy, motor neuron disease, and chronic progressive external ophthalmoplegia. These limitations preclude the use of the current model in clinical practice.

For monitoring purposes, automated classification of disease severity that could be used by patients themselves in addition to standard care, could potentially lead to an improvement of clinical care. It could reduce the need to travel long distances for control visits, assist patients in adapting the dose of their maintenance medication according to prespecified personalized rules and potentially avoid hospitalization by immediately starting emergency treatment in case of an alarming deterioration. In this exploratory study has demonstrated that facial weakness is quantifiable in videos using facial expression recognition software. In addition, the results of the DL model show that binary classification of disease severity is possible with high accuracy. Our results suggest that the degree of facial movements, as a reflection of facial weakness, is a good predictor for overall muscular weakness. This supports the use of facial weakness to assess disease severity in MG. However, two important limitations remain: our results were obtained on a dichotomous classification instead of the continuous QMG score. Furthermore, these results should be validated in a longitudinal study to investigate the ability to detect changes within individuals.

Limitations

Identifying the elements in the data used by convolutional neural networks for classification is not straightforward. Although techniques such as class activation mapping have been developed to identify areas of interest in still images,[40] class activation mapping algorithms for video data were not available at the time of these experiments. However, as all external factors (lighting, green screen, camera angle) were standardized, classification could not have been based on anything other than the face and its movement. In addition, automated analyses with FaceReader demonstrated quantifiable differences in facial expressions on which a deep learning could be trained.

Unexpectedly, the performance of the current DL model was better on the unseen validation set than on the original training set. This is a surprising finding because the purpose of the additional validation set was to test whether training results could be reproduced with unseen data, not to further improve the model. There are two possible explanations for this finding. The QMG score of patients with moderate–severe disease was significantly higher in the dataset used for validation compared to the training set. This potentially made it more easy to detect a difference in disease severity or to discriminate patients from HC in this dataset. Alternatively, the observed higher diagnostic yield may have been a spurious result caused by the relatively small sample size of the validation set. Given the rarity of MG, this was the maximum number of patients we were able to include within a reasonable time frame.

Finally, because of the exploratory nature and relatively small sample size of this study, no covariates were used in the analysis. However, several covariates that affect the performance of facial recognition algorithms have been identified.[41] These covariates include age, sex, and ethnicity. In short, older people are easier to recognize by facial recognition software than younger people[41]. There is also an effect of sex on the accuracy of facial recognition software but there is disagreement whether men or women are more easily recognized, and this effect decreases with increasing age.[41] Ethnicity can affect the performance of facial recognition software as ethnic differences in facial characteristics may lead to different results. Furthermore, an algorithm trained on a specific ethnic group may yield incorrect results when applied to people of a different background.[41] Participants included in this study were a reflection of the Dutch population and therefore mostly of European descent, although this was not formally assessed. Therefore, caution is advised when extrapolating these results to patients groups of other ethnicities.

Due to the exploratory nature of this study, we did not apply a post-hoc correction. A post hoc Bonferroni correction on the quantification of facial weakness (FaceReader) would have resulted in significant levels of 0.0004 for the AUs and 0.008 for the six emotions.

Author Contributions

Annabel Ruiter, Jan Verschuuren, and Martijn Tannemaat contributed to the study concept and design. Annabel Ruiter, Ziqi Wang, Zhao Yin, Willemijn Naber, Jerrel Simons, Jan van Gemert, Jan Verschuuren, and Martijn Tannemaat contributed to the acquisition and analysis of data. Annabel Ruiter, Ziqi Wang, Jurre Blom, Jan van Gemert, Jan Verschuuren, and Martijn Tannemaat contributed to drafting the text and/or figures.

Acknowledgments

Dr. Erik van Zwet, statistician, department of biostatistics LUMC, was consulted to aid in the development of specific parts of the study protocol and during the analysis of collected data. We thank Jeanette Wigbers, our outpatient nurse, for performing the QMG tests for all

participating patients. Funding for this project originated from the LUMC Neuromuscular Fund, Target2B!, and it is part of the research programme C2D–Horizontal Data Science for Evolving Content with project name DAC-COMPLI and project number 628.011.002, which is (partly) financed by the Netherlands Organization for Scientific Research (NWO) and is also supported by the Leiden Institute of Advanced Computer Science (LIACS).

Conflicts of Interest

A. M. Ruiter, W. C. Naber, J. Simons, Z. Wang, Z. Yin, and J. C. van Gemert, report no disclosures relevant to the manuscript. M. R. Tannemaat has been involved in MG research sponsored by Argenx, Alexion, and NMD Pharma. All reimbursements were received by the LUMC, M. R. Tannemaat had no personal financial benefit from these activities. J. J. G. M. Verschuuren receives financial support from Target to B consortium, Prinses Beatrix Spierfonds, and has been involved in trials or consultancies for Argenx, Alexion, and Rapharma. He is a coinventor on patent applications based on MuSK-related research. The LUMC received royalties from IBL and Argenx for MG research. All reimbursements were received by the LUMC. Several authors of this publication are members of the Netherlands Neuromuscular Center and the European Reference Network for rare neuromuscular diseases EURO-NMD.

Data Availability Statement

Raw quantitative data, including anonymized individual participants' data, that support the findings of this study are available from the corresponding author, upon reasonable request. Privacy laws preclude the sharing of facial videos of individual study participants.

References

1. Verschuuren JJ, Palace J, Gilhus NE. Clinical aspects of myasthenia explained. Autoimmunity. 2010;43(5–6):344-352.
2. Keesey JC. Clinical evaluation and management of myasthenia gravis. Muscle Nerve. 2004;29(4):484-505.
3. de Meel RHP, Tannemaat MR, Verschuuren J. Heterogeneity and shifts in distribution of muscle weakness in myasthenia gravis. Neuromuscul Disord. 2019;29(9):664-670.
4. Gilhus NE, Tzartos S, Evoli A, Palace J, Burns TM, Verschuuren J. Myasthenia gravis. Nat Rev Dis Primers. 2019;5(1):30.
5. Valko Y, Rosengren SM, Jung HH, Straumann D, Landau K, Weber KP. Ocular vestibular evoked myogenic potentials as a test for myasthenia gravis. Neurology. 2016;86(7):660-668.
6. Jani-Acsadi A, Lisak RP. Myasthenic crisis: guidelines for prevention and treatment. J Neurol Sci. 2007;261(1–2):127-133.
7. Steels J. Improving outcomes by tracking symptoms, triggers and quality of life. Real-World Assessment and Patient Perceptions of a Prototype Myasthenia Gravis Tracking App Oral Presentation MGFA Scientific Session. AANEM; 2022.
8. Haselkorn A, Coye MJ, Doarn CR. The future of remote health services: summary of an expert panel discussion. Telemed J E Health. 2007;13(3):341-347.
9. Sapci AH, Sapci HA. Innovative assisted living tools, remote monitoring technologies, artificial intelligence-driven solutions, and robotic Systems for Aging Societies: systematic review. JMIR Aging. 2019;2(2):e15429.
10. González-Navarro FF, Belanche-Muñoz LA, Silva-Colón KA. Effective classification and gene expression profiling for the facioscapulohumeral muscular dystrophy. PLoS One. 2013;8(12):e82071.
11. González-Navarro FF, Belanche-Muñoz LA, Gámez-Moreno MG, Flores-Ríos BL, Ibarra-Esquer JE, López-Morteo GA. Gene discovery for facioscapulohumeral muscular dystrophy by machine learning techniques. Genes Genet Syst. 2016;90(6):343-356.
12. Morrow JM, Sormani MP. Machine learning outperforms human experts in MRI pattern analysis of muscular dystrophies. Neurology. 2020;94(10):421-422.
13. Felisaz PF, Colelli G, Ballante E, et al. Texture analysis and machine learning to predict water T2 and fat fraction from non-quantitative MRI of thigh muscles in facioscapulohumeral muscular dystrophy. Eur J Radiol. 2021;134:109460.
14. Wang K, Romm EL, Kouznetsova VL, Tsigelny IF. Prediction of premature termination codon suppressing compounds for treatment of Duchenne muscular dystrophy using machine learning. Molecules. 2020;25(17):3886.
15. Boonipat T, Asaad M, Lin J, Glass GE, Mardini S, Stotland M. Using artificial intelligence to measure facial expression following facial reanimation surgery. Plast Reconstr Surg. 2020;146(5):1147-1150.
16. Mothes O, Modersohn L, Volk GF, et al. Automated objective and marker-free facial grading using photographs of patients with facial palsy. Eur Arch Otorhinolaryngol. 2019;276(12):3335-3343.
17. Liu X, Xia Y, Yu H, Dong J, Jian M, Pham TD. Region based parallel hierarchy convolutional neural network for automatic facial nerve paralysis evaluation. IEEE Trans Neural Syst Rehabil Eng. 2020;28(10):2325-2332.
18. Ding M, Kang Y, Yuan Z, Shan X, Cai Z. Detection of facial landmarks by a convolutional neural network in

patients with oral and maxillofacial disease. Int J Oral Maxillofac Surg. 2021;50:1443-1449.
19. Lee SA, Kim J, Lee JM, Hong YJ, Kim IJ, Lee JD. Automatic facial recognition system assisted-facial asymmetry scale using facial landmarks. Otol Neurotol. 2020;41(8):1140-1148.
20. Ten Harkel TC, Speksnijder CM, van der Heijden F, Beurskens CHG, Ingels K, Maal TJJ. Depth accuracy of the RealSense F200: low-cost 4D facial imaging. Sci Rep. 2017;7(1):16263.
21. Ruiter AM, Naber WC, Verschuuren J, Tannemaat MR. The face of myasthenia gravis. Neurology. 2020;95 (2):89-90.
22. Katzberg HD, Barnett C, Merkies IS, Bril V. Minimal clinically important difference in myasthenia gravis: outcomes from a randomized trial. Muscle Nerve. 2014;49 (5):661-665.
23. https://www.noldus.com/facereader/facial-expression-analysis
24. Ekman P, Friesen WV, Hager JC. Facial Action Coding System: Facial Action Coding System: The Manual: On CD-ROM. Research Nexus; 2002.
25. Soomro K, Zamir AR, Shah M. Ucf101: a dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:12120402 2012.
26. Kay W, Carreira J, Simonyan K, et al. The kinetics human action video dataset. arXiv preprint arXiv:170506950 2017.
27. https://github.com/gsig/PyVideoResearch
28. Carreira J, Zisserman A. Quo vadis, action recognition? A new model and the kinetics dataset. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2017:6299-6308.
29. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2016:770-778.
30. Anand K, Wang Z, Loog M, van Gemert J. Black magic in deep learning: how human skill impacts network training. arXiv preprint arXiv:200805981 2020.
31. Hutter F, Hoos H, Leyton-Brown K. An efficient approach for assessing hyperparameter importance. International Conference on Machine Learning; 2014:754-762.
32. Koutsoukas A, Monaghan KJ, Li X, Huan J. Deep-learning: investigating deep neural networks hyper-parameters and comparison of performance to shallow methods for modeling bioactivity data. J Chem. 2017;9 (1):42.
33. Coulson SE, O'Dwyer NJ, Adams RD, Croxson GR. Expression of emotion and quality of life after facial nerve paralysis. Otol Neurotol. 2004;25(6):1014-1019.
34. van de Geest-Buit WA, Rasing NB, Mul K, et al. Facing facial weakness: psychosocial outcomes of facial weakness and reduced facial function in facioscapulohumeral muscular dystrophy. Disabil Rehabil. 2022;45:1-10.
35. de Meel RHP, Raadsheer WF, van Zwet EW, Tannemaat MR, Verschuuren J. Ocular weakness in myasthenia gravis: changes in affected muscles are a distinct clinical feature. J Neuromuscul Dis. 2019;6(3):369-376.
36. Zhuang Y, McDonald M, Uribe O, et al. Facial weakness analysis and quantification of static images. IEEE J Biomed Health Inform. 2020;24(8):2260-2267.
37. Latorre-Pellicer A, Ascaso Á, Trujillano L, et al. Evaluating Face2Gene as a tool to identify Cornelia de Lange syndrome by facial phenotypes. Int J Mol Sci. 2020;21 (3):1042.
38. Gurovich Y, Hanani Y, Bar O, et al. Identifying facial phenotypes of genetic disorders using deep learning. Nat Med. 2019;25(1):60-64.
39. Hadj-Rabia S, Schneider H, Navarro E, et al. Automatic recognition of the XLHED phenotype from facial images. Am J Med Genet A. 2017;173(9):2408-2414.
40. Hiley L, Preece A, Hicks Y, Chakraborty S, Gurram P, Tomsett R. Explaining motion relevance for activity recognition in video deep learning models. arXiv preprint arXiv:200314285 2020.
41. Lui Y, Bolme D, Draper B, Beveridge J, Givens G, Phillips PJ. A meta-analysis of face recognition covariates. IEEE 3rd International Conference on Biometrics: Theory, Applications, and Systems; 2009:1-8.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Table S1.

