Artificial intelligence-based classification of motor unit action potentials in real-world needle EMG recordings

Hubers, D.; Potters, W.; Paalvast, O.; Jonge, S. de; Doelkahar, B.; Tannemaat, M.; ... ; Verhamme, C.

Clinical Neurophysiology 156 (2023) 220–227

Contents lists available at ScienceDirect

Clinical Neurophysiology

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

Artificial intelligence-based classification of motor unit action potentials in real-world needle EMG recordings

Deborah Hubers [a,*], Wouter Potters [a], Olivier Paalvast [a], Sterre de Jonge [a], Brian Doelkahar [a], Martijn Tannemaat [b], Luuk Wieske [a,c,1], Camiel Verhamme [a,1]

[a] Department of Neurology and Clinical Neurophysiology, Amsterdam Neuroscience, Amsterdam University Medical Centers, Location AMC, Amsterdam, the Netherlands
[b] Department of Neurology, Leiden University Medical Center, Leiden, the Netherlands
[c] Department of Clinical Neurophysiology, St. Antonius Hospital, Nieuwegein, the Netherlands

HIGHLIGHTS

- Two nested artificial neural networks can classify MUAP duration in real world needle EMG recordings obtained during routine care.
- A first model accurately classified segments as rest, contraction or artifacts.
- A second model subsequently classified MUAP duration in contraction segments with moderate accuracy.

ARTICLE INFO

Article history:
Accepted 17 October 2023
Available online 3 November 2023

Keywords:
Needle EMG
Motor unit action potentials
Artificial intelligence
Diagnostic accuracy

ABSTRACT

Objective: To develop an artificial neural network (ANN) for classification of motor unit action potential (MUAP) duration in real-word, unselected and uncleaned needle electromyography (n-EMG) recordings.
Methods: Two nested ANN models were trained, the first discerning muscle rest, contraction and artifacts in n-EMG recordings from 2674 individual muscles from 326 patients obtained as part of daily care. The second ANN model subsequently used segments labeled as contraction for prediction of prolonged, normal and shortened MUAPs. Model performance was assessed in one internal and two external validation datasets of 184, 30 and 50 muscles, respectively.
Results: The first model discerned rest, contraction and artifacts with an accuracy of 96%. The second model predicted prolonged, normal and shortened MUAPs with an accuracy of 67%, 83% and 68% in the different validation sets.
Conclusions: We developed a two-step ANN that classifies rest, muscle contraction and artifacts from real-world n-EMG recordings with very high accuracy. MUAP duration classification had moderate accuracy.
Significance: This is the first study to show that an ANN can classify MUAPs in real-world n-EMG recordings highlighting the potential for AI assisted MUAP classification as a clinical tool.

1. Introduction

Classification of motor unit action potentials (MUAPs), obtained by needle electromyography (n-EMG), is mostly done by clinicians at the bedside through auditory and visual assessment of MUAP characteristics, like duration, amplitude and number of phases (Preston and Shapiro, 2013). In this assessment, changes in MUAP duration are thought to be most specific and indicative of a neurogenic or myopathic disorder (Preston and Shapiro, 2013). Human ear can precisely and quickly recognize (subtle) pathologies in MUAP duration, as supported by a simulation study that showed that humans could hear differences in the duration of simulated MUAPs of 1 ms (Katirji, 2014). However, this method of MUAP classification is dependent on clinician's training and exposure. Kendall et al studied six n-EMG cases reviewed by 66 blinded examiners and found that overall diagnostic agreement was only 50 %, although it increased with experience (Kendall and Werner, 2006). Low interrater agreement and accuracy of MUAP classification, especially in the setting of clinicians not routinely performing

* Corresponding author at: Department of Neurology and Clinical Neurophysiology, Amsterdam Neuroscience, Amsterdam University Medical Centers, Location AMC, Amsterdam, the Netherlands.
E-mail address: d.hubers@amsterdamumc.nl (D. Hubers).
[1] These authors contributed equally.

https://doi.org/10.1016/j.clinph.2023.10.008
1388-2457/

n-EMG, may lead to misdiagnosis or diagnostic delay (Kendall and Werner, 2006).

Quantitative methods can provide numerical estimates of MUAP characteristics which can improve inter-rater agreement of MUAP classification. However, MUAP quantification is time-consuming compared to bedside auditory and visual assessment by clinicians (Fuglsang-Frederiksen, 2006; Farkas et al., 2010). Artificial intelligence (AI) has been studied as an alternative objective classification method for MUAPs. An AI algorithm may be trained to interpret MUAP characteristics and recruitment patterns during contraction and (abnormal) spontaneous activity during rest, while it may also rate the quality of the signal, making AI a more flexible and complete tool than EMG quantification. So far, AI studies reported high accuracy for some of these aspects, such as 100 % accuracy for classification of resting n-EMG signals and 90–98 % for MUAP classification (Katsis et al., 2007; Chatterjee et al., 2020; Mokdad et al., 2020; Subasi, 2012; Christodoulou and Pattichis, 1999). However, these studies were all limited by the use of pre-selected n-EMG segments from which artifacts were removed manually, whereas real-world n-EMG, recorded as part of daily care, contain many artifacts and non-informative segments. Moreover, limited samples were studied, ranging from 18-62 patients, and no studies used an external validation cohort. Therefore, generalization and potential clinical implementation is limited so far.

AI algorithms could be developed further in two ways. First, AI may be trained using clinical diagnoses as the basis. For example, a recent study using amyotrophic lateral sclerosis and inclusion body myositis showed that a machine learning algorithm could discriminate between these disorders without any input on actual MUAP characteristics (Tannemaat et al., 2023). This approach allows for unsupervised learning, transcending clinical paradigms, but may be limited because disease specific abnormalities may not be readily translatable to myopathic or neuropathic disorders in general. In addition, depending on the disorder, muscles studied with n-EMG may be severely, slightly or not affected, complicating how an overall diagnosis on a patient-level should be translated to abnormalities on the muscle-level. Alternatively, AI may be trained to classify known MUAP characteristics that are used in the clinic such as duration. AI predictions on these characteristics can subsequently be incorporated by the clinician into an overall conclusion on the most likely electrophysiological diagnosis. This approach may more easily complement current clinical practice and may therefore be easier to incorporate in routine clinical practice.

The goal of this study was to design and validate an artificial neural network (ANN) for MUAP duration classification using unselected and uncleaned real-world n-EMG data obtained from clinical practice.

2. Methods

2.1. Design and ethics

For this retrospective study, we set up a data flow of two consecutive ANNs mimicking clinical practice. The first ANN (the "muscle activity classification model") was designed to label each data point in a complete n-EMG recording as an assessment of activity at rest, contraction or needle movement or other artifacts. The second ANN (the "MUAP duration classification model") subsequently used only those n-EMG segments labeled as contraction and was designed to label MUAP duration in these segments as prolonged, normal or shortened. The medical ethical committee waived the need for informed consent because of the use of anonymized data that was collected as part of clinical care. Participants were offered an opt-out option by mail to ascertain if they objected to the use of their data for scientific studies. This paper was drafted according to the Standards for Reporting Diagnostic accuracy studies (STARD) guidelines on reporting of studies of diagnostic accuracy (Cohen et al., 2016).

2.2. Participants and settings

Data for model development were collected as part of clinical care in the department of Clinical Neurophysiology of the Amsterdam University Medical Centers (Amsterdam UMC), location AMC, Amsterdam, the Netherlands a tertiary referral center for neuromuscular diseases. Between 29 January 2020 and 22 April 2020, all n-EMG examinations were consecutively recorded and used for this study regardless of the reason for study or clinical diagnosis. This consecutive cohort was enriched with n-EMG recordings stored between December 2016 up and November 2021 of patients suspected of myositis to add short duration MUAPs. We excluded bulbar muscles (e.g. tongue base, sternocleidomastoid or masseter muscles) because of the different MUAP characteristics, n-EMG recordings for which the reference standard had not been stored, and n-EMG recordings labeled as having both prolonged and shortened MUAPs in the same muscle (like what can be observed in chronic myopathies) (Subasi, 2012). For the MUAP duration classification model, n-EMG recordings that did not contain contraction segments were also excluded.

The Amsterdam UMC dataset was split into a training, test and validation set. Next to this validation set, two additional external validation cohorts were used. First, a non-consecutive convenience sample from patients with n-EMG recordings collected at the department of Clinical Neurophysiology, Leiden UMC, the Netherlands (LUMC dataset) that had been classified as either normal or neuropathic. Second, we used the n-EMGs recordings from the online available EMGLab database (Nikolic, 2001) classified as neuropathic, normal or myopathic. For both validation sets, we assumed that neuropathic recordings mainly contained prolonged MUAPs and myopathic recordings contained shortened MUAPs.

2.3. Data collection

Muscles examined by n-EMG were selected based on the clinical question, as judged by the examiner, and clinical protocols. Each n-EMG recording involved assessment of activity at rest and during voluntary contraction. In general, n-EMG assessment took place in three phases: with the muscle at rest, during slight contraction and subsequently gradually increasing to at least moderate contraction. The examiner optimized signal quality throughout the recording by re-positioning the needle during the exam so to obtain a reliable recording of one or more MUAPs firing. All signals were recorded and exported anonymously as audio files (uncompressed Wav files). For this study, complete and uncleaned n-EMG recordings were used that were recorded as part of routine clinical care and without any preprocessing steps other than regular filtering during acquisition. Filter settings during acquisition for the Amsterdam UMC and the LUMC were: a low-pass filter of 10 kHz, a high-pass filter of 10 Hz. Filter settings during acquisition for the EMGLab database were: a low-pass filter of 10 kHz, a high-pass filter of 2 Hz. The sampling frequency of the audio files from the Amsterdam UMC and the LUMC was 44100/s, and the length varied from file to file (2 s up to 120 s). The sampling frequency of the audio files from the EMGLab database was 23437.5/s.

2.4. Data preprocessing

Data preprocessing is described in detail in the Supplementary Methods. In short, for the muscle activity classification model, n-EMG recordings were normalized, split into 2 s segments using a

sliding window approach of 0.1 s creating overlapping segments. For the MUAP classification model, n-EMG recordings were normalized and split into 2 s segments creating consecutive segments. For the muscle activity classification model, all 2 s segments were used in the following steps. For the MUAP duration classification model, only 2 s segments classified as contraction were used. From the 2 s segments, spectrograms were created and scaled to the Mel scale (Sahidullah and Saha, 2012; Rao and Manjunath, 2017). The Mel scale approximates the human auditory system's response more closely than the linearly spaced frequency bands used in the normal spectrum (Sahidullah and Saha, 2012; Umesh et al., 1999). For the muscle activity classification model, we included frequencies 0–10000 Hz. For the MUAP duration classification model, we included frequencies 500–5000 Hz because frequencies below 500 Hz represent distant MUAPs and frequencies above 5000 Hz noise (Tankisi et al., 2020).

2.5. ANN model development: Model 1 muscle activity classification model

Details on development of the muscle activity classification model are described in the Supplementary Methods. As reference standard, muscle activity was annotated throughout each n-EMG recording as four class labels: i.e. muscle rest, voluntary contraction, needle movement artifacts, or non-analyzable (e.g. technical artifacts) by the investigators. This was done by first developing in-house annotation software (Python 3.9) with a user interface to annotate a complete n-EMG recording based on auditory and visual inspection. The four class labels were continuously assigned to 95n-EMG recordings of 120 s each by at least two investigators [CV, LW, WVP, DH]. If unsure, an investigator set no annotation. Class labels for each data point were combined and labels were maintained only in case of majority agreement of at least two investigators. Subsequently, two-second segments were labeled as contraction (in case all data points in that segment had been labeled as contraction by both investigators), as rest (in case all data points in that segment had been labeled as rest by both investigators) or as artifact (needle movement/non-analyzable combined; in case $\geq$15 % ($\geq$0.3 s for 2 s) of data points in that segment had received this label). Nine 120 s n-EMG recordings were annotated by all four examiners to determine the inter-rater reliability by calculating Krippendorff's alpha (Zapf et al., 2016).

Next, we applied transfer learning to train the InceptionResNetV2 model, a convolutional neural network (CNN) pre-trained on ImageNet (Szegedy et al., 2017), for classification of the 3 labels. To increase the data segments and equalize distributions for each reference standard category, we augmented the data by skewing and distorting the Mel spectrograms (Nodera et al., 2019). We first trained the CNN using 4-fold cross-validation with 90 % from the annotated segments. The remaining 10 % segments was then used as an internal validation set. There was no overlap between muscles included in the training, test and validation set. Confusion matrices were created and the model's precision, recall, F1-score, and accuracy were on the validation sets. Based on the test loss, the best model was evaluated in the validation set and used to select data segments for the next model, the MUAP duration classification model.

2.6. ANN model development: Model 2 MUAP duration classification model

Details on development of the MUAP duration classification model are described in the Supplementary Methods. As reference standard, we used the original MUAP duration classification as determined by experienced neurophysiologists at the bedside as part of clinical care. MUAPs of different durations may be present in a n-EMG recording, therefore MUAPs could be classified as: prolonged duration, normal duration, shortened duration, a mix of prolonged and shortened duration, borderline (between prolonged and normal duration), and a mix of normal and shortened duration. For each complete n-EMG recording, one overall label was provided at the bedside that best reflected the recording of that muscle. We reclassified borderline as normal and shortened and normal duration as shortened duration and excluded recordings that had both shortened and prolonged MUAPs. This led to three labels that were used for model training: prolonged duration, normal duration, and shortened duration.

Next, we again applied transfer learning to train the InceptionResNetV2 model for MUAP duration classification. Only two-second segments predicted as contraction by the muscle activity classification model (see above) were included in this model. All the two-second segments predicted as contraction were used regardless of the level of contraction so that segments containing individually discernable MUAPs up to segments containing interference patterns were included in the model. N-EMG recordings of individual muscles containing more than 10 two-second segments predicted as contraction were included and divided into a training and testing set (90 %) and a validation set (10 %), stratified to obtain comparable distributions of the 3 MUAP duration classes. To increase the number of data segments and equalize distributions for each reference standard category, we augmented the data by skewing and distorting the Mel spectrograms (Nodera et al., 2019). There was no overlap between muscles included in the training, test and validation set. The MUAP classification model was trained and tested using 4-fold cross-validation on 90 % of the data set. Model output was determined by calculating the proportion of two-second segments per recording where MUAP duration was predicted as prolonged, normal or shortened. The label with the highest proportion was used by the model as an overall label for the entire recording. Confusion matrices for overall labels and the reference standard were created and the best model was chosen based on test loss. The accuracy of the best performing model was subsequently validated with the internal 10 % validation set from the Amsterdam UMC, the external validation dataset from the LUMC and the EMGLab database.

2.7. Secondary analyses: Power threshold for clinical MUAP interpretation

In clinical practice, MUAP characteristics are usually assessed during slight contraction as this allows for inspection of individual MUAPs. To mimic this, we retrained the MUAP duration classification model using only contraction segments in which individual MUAP characteristics could be reliably interpreted as would be done clinically. This was defined using a cut-off based on the power of that segment. The power of a signal is the sum of the absolute squares of its time-domain samples divided by the signal length (Meinsma et al., 2019). Two interpreters [CV, LW] independently scored 62 two-second randomly selected contraction-labeled segments of varying power as interpretable or not. The maximum power at which both interpreters had scored a segment to be interpretable was used as cut-off.

2.8. Secondary analyses: MUAP classification model compared to human performance

We investigated the clinical utility of the MUAP duration classification model by comparing its diagnostic yield to the performance of three experienced clinical neurophysiologists [MT, LW and CV] who classified the same EMGLab data n-EMG recordings as having MUAPs of prolonged, normal, or shortened duration in

the same manner that they would do during routine clinical assessments. N-EMG recordings were assessed independently and in random order and the raters did not know the frequencies of abnormalities. None of the investigators had seen this dataset before. Diagnostic yield of each human interpreter was calculated and compared to performance of the MUAP duration classification model.

2.9. Data availability

Data not provided in the article may be shared (anonymized) at the request of any qualified investigator for purposes of replicating procedures and results.

3. Results

The Amsterdam UMC data set for model development consisted of n-EMG recordings from 2671 individual muscles from 326 patients.

3.1. ANN model 1: muscle activity classification model

For development and training of the muscle activity classification model, the 95 annotated n-EMG files were split in 85 n-EMG recordings to train and test the model, while the remaining 10 files were used for validation. On average, 24.8 % of each recording was excluded due to lack of agreement between examiners. The inter-rater reliability of the different investigators was 0.77 (Krippendorff alpha). Extraction of the two-second segments resulted in 64,994 segments in the training set and 6417 segments in the validation set. Table 1 shows the confusion matrix of the best muscle activity classification model evaluated on the validation data set. Accuracy of this model on the validation data set was 96 %. Precision, recall and F1-scores for contraction were 99 %, 96 % and 97 % respectively (for other labels, see Table 2).

3.2. ANN model 2: MUAP duration classification

For development and training of the MUAP classification model, 2341 from the 2674 n-EMG recordings were used (163 recordings excluded because of bulbar muscle, 115 were excluded because MUAP duration assessment had not been stored, and 55 recordings had both prolonged and shortened MUAPs). From these n-EMG recordings, 2108 (90 %) n-EMG recordings were used for training and testing, and 233 (10 %) were used as a validation set. This division was stratified based on class labels of MUAP duration. After running the muscle activity classification model and selecting contraction segments, an additional 49 n-EMG recordings were excluded because no contraction segments were present. The external validation set from the LUMC consisted of 30 muscles (17 prolonged and 13 normal). This dataset did not contain muscles with MUAPs of shortened duration. The EMGLab database consisted of 50 muscles (19 shortened, 21 prolonged, and 10 normal).

Fig. 1 shows proportions of two-second segments of each n-EMG recording predicted by the best MUAP classification model as prolonged, normal and shortened stratified according to the reference standard in the Amsterdam UMC validation dataset. For most muscles with prolonged MUAPs according to the reference standard, indeed most segments were predicted as prolonged. However, segments were predicted as prolonged in a considerable number of muscles with normal MUAPs. For muscles with shortened MUAPs, most segments were indeed predicted as shortened.

Table 3 shows the confusion matrix of the best MUAP classification model evaluated on the different validation datasets. Accuracy was 67 % for the Amsterdam UMC validation dataset, 83 % for the LUMC dataset and 68 % for the EMGLab dataset. Table 4 shows the performance matrices for the different validation datasets. Within the different validation data sets, two important misclassification patterns were observed: 1) normal MUAPs that were predicted to be prolonged (observed in the Amsterdam UMC and LUMC validation datasets) and 2) normal MUAPs that were predicted to be shortened (or vice versa; observed in the Amsterdam UMC validation data set and the EMGLab database).

3.3. Power threshold for clinical MUAP interpretation

After evaluation by two investigators, the maximum power of a segment at which MUAPs could be interpretable as would be done in the clinic was $1.26 * 10^8$ (V^2/Hz). Extraction of the 2 s segments containing contraction with a power below this cut-off resulted in a total of 11,753 segments in the training and testing set (prolonged 2722, normal 5215, shortened 3816). Training of the MUAP classification model using only two-second segments with a maximum power below this cut-off decreased model accuracy and per-

Table 1
Confusion matrix ANN model 1: muscle activity classification model. Distribution of correct and incorrect predictions of the muscle activity classification model: an artificial neural network (ANN) model developed to label segments of needle-EMG recordings as rest, voluntary contraction or artifacts (i.e. needle movement or non-analyzable segments). Data shown are the results for 10 needle-EMG recordings with 6417 two s segments in the validation set (not used during training). As a reference standard, labels about which two investigators agreed were used (see methods). The shaded cells indicate the correctly classified segments.

		ANN model 1 muscle activity classification model prediction		
		Rest	Contraction	Artifact
Reference standard	Rest	1502	18	63
	Contraction	97	2386	5
	Artifact	27	15	2304

Table 2
Performance ANN model 1: muscle activity classification model. Performance of the muscle activity classification model in terms of precision, recall and F1-score for all rest, contraction and artifact as well as overall accuracy. Data shown are based on 10 needle-EMG recordings with 6417 two-second segments in the validation dataset. The *support* column indicates the number of two-second segments used for each measure.

	Precision	Recall	F1-score	Accuracy	Support
Rest	92 %	95 %	94 %		1883
Contraction	99 %	96 %	97 %		2488
Artifact	97 %	98 %	98 %		2346
Overall				96 %	6417

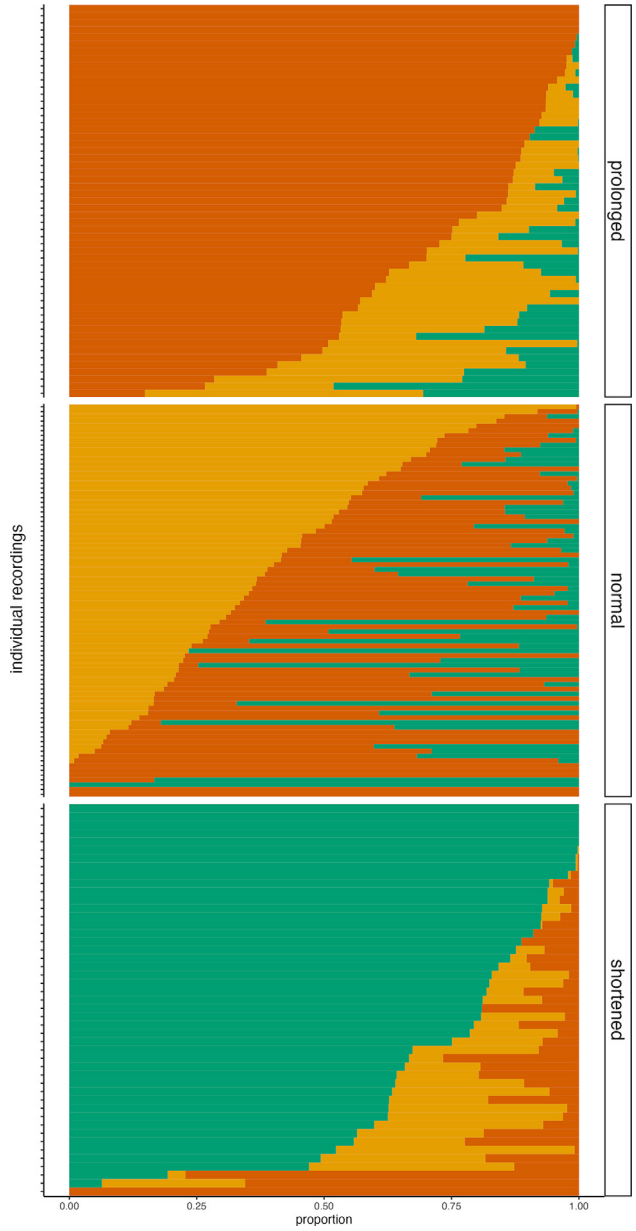

Fig. 1. Distribution of predicted segments according to the reference standard. Figure displaying proportions of two s segments predicted as prolonged, normal and shortened in the Amsterdam University Medical Centers (Amsterdam UMC) validation set consisting of 233 needle EMG (n-EMG) recordings of 120 s, according to the reference standard for that recording. Each horizontal line represents 1 n-EMG recording in 1 muscle. Red indicates segments predicted as having motor unit action potentials (MUAPs) with a prolonged duration, yellow as normal and green as shortened duration.

formance to an accuracy of 46 % in the Amsterdam UMC dataset (see Supplementary Table 4 for the performance matrix).

3.4. Human performance of MUAP classification

Three examiners (CV, LW, MT) annotated all 50 n-EMG recordings in the EMGLab database. Accuracy was 62 % for all three different investigators. Supplementary Table 5 shows the confusion matrices for the different examiners.

4. Discussion

We developed a two-step ANN to classify MUAP duration mimicking daily clinical practice. The first model discerned rest, muscle contraction and artifacts from real-world n-EMG recordings with very high accuracy. The second model subsequently used muscle contraction segments to classify MUAP duration in these segments as prolonged, normal or shortened with moderate accuracy.

This two-step approach illustrates the possibilities to use ANNs as an adjunctive diagnostic tool in a real-world n-EMG setting. This is the first study that has used AI in unprocessed real-world n-EMG recordings and the first study to develop an ANN to discern rest, muscle contraction or artifacts using n-EMG recordings. Previous studies on studying detection of muscle contraction using AI have been limited to surface EMG recordings which are not commonly used for diagnostic investigations. The first-step ANN model we trained to classify rest, muscle contraction or artifacts in n-EMG recordings showed very high accuracy overall. Performance was least for the prediction of muscle rest, although performance indices were still $\geq$92 %, and the most frequent misclassification was the classification of segments containing contraction as rest.

Model performance of our second step MUAP duration classification ANN was moderate whereas other studies reported much better performances with accuracies approaching 100 % for various tasks (Bakiya et al., 2020; Chatterjee et al., 2020; Doulah and Fattah, 2014; Mishra et al., 2016; Mokdad et al., 2020; Nagineni et al., 2018; Samanta et al., 2020; Sengur et al., 2017). However, important limitations of previous studies are the use of manually selected n-EMG data which does not reflect daily clinical care, the limited sample sizes and the lack of independent data validation sets which is likely to inflate model performance. As an example, most previous studies used the EMGLab dataset, which we used as an external validation dataset, to train and test their model without external validation (Bakiya et al., 2020; Chatterjee et al., 2020; Doulah and Fattah, 2014; Mishra et al., 2016; Mokdad et al., 2020; Nagineni et al., 2018; Samanta et al., 2020; Sengur et al., 2017). The EMGLab database is a research dataset in which n-EMG signals were recorded at a specific level of contraction at predefined locations only when the signal quality was sufficient. The n-EMG recordings in this dataset therefore represent an ideal test set which varies greatly from n-EMG recordings obtained during routine clinical care.

Despite the lower accuracy of our model based on real world n-EMG recordings, the model still shows important potential. No major misclassification occurred between shortened or prolonged MUAPs, which is reassuring as this resembles the clinical separation of two characteristics at opposing ends of a spectrum. Interestingly, model performance did not improve when only using segments of slight contraction while, for clinical evaluation, slight levels of contraction are preferred as individual MUAPs can still be discernible. This indicates that the model can also extract relevant information from segments with higher levels of contraction where human interpretation fails. Future work may provide more insight into the data features or segments used by the ANN model for classification. Finally, an ANN model lacks important clinical information available to a clinician when performing n-EMG recordings, such as the apparent level of contraction, presence of atrophy or which segments of the recording are most reliable. When we compared human interpreters to our MUAP duration classification model, both operating in the same setting of only having access to the raw n-EMG signal, we found highly similar accuracies.

Table 3
Confusion matrix ANN model 2: MUAP duration classification. Table showing the distribution of correct and incorrect predictions of the second artificial neural network (ANN) model developed to classify motor unit action potentials (MUAP) duration as prolonged, normal, or shortened. Table shows results as obtained for all three validation datasets (see methods). The ANN MUAP duration classification model labeled all two-second segments in a needle-EMG recording that were predicted as contraction by the first ANN model (see methods). The frequency of each label in the overall recording was calculated and the label with the highest proportion was used as an overall label for the entire recording. The shaded cells indicate the correctly classified segments.

		ANN model 2: MUAP duration classification		
		Prolonged	Normal	Shortened
Amsterdam UMC validation set				
Reference standard	Prolonged	51	3	1
	Normal	43	28	11
	Shortened	3	0	44
LUMC validation set*				
Reference standard	Prolonged	17	0	
	Normal	5	8	
EMGlab validation set				
Reference standard	Prolonged	18	3	0
	Normal	0	5	5
	Shortened	3	5	11

Amsterdam UMC: Amsterdam University Medical Centers; LUMC: Leiden University Medical Center.
* No MUAPs with shortened duration available in this data set.

Table 4
Performance ANN model 2: MUAP duration classification. Performance of the second artificial neural network (ANN) model developed to classify motor unit action potential (MUAP) duration as prolonged, normal or shortened that had performed best during training, based on validation loss, and presented for the different validation datasets (see methods). Support indicates the number of needle-EMG recordings used for each measure.

	Precision	Recall	F1-score	Accuracy	Support
Amsterdam UMC validation set					
Prolonged	53 %	93 %	67 %		55
Normal	90 %	34 %	50 %		82
Shortened	79 %	94 %	85 %		47
Overall				67 %	184
LUMC validation set*					
Prolonged	81 %	100 %	90 %		17
Normal	100 %	62 %	77 %		13
Overall				83 %	30
EMGlab validation set					
Prolonged	86 %	86 %	86 %		21
Normal	38 %	50 %	43 %		10
Shortened	69 %	58 %	63 %		16
Overall				68 %	50

Amsterdam UMC: Amsterdam University Medical Center; LUMC: Leiden University Medical Center.
* No MUAPs with shortened duration available in this data set.

The most common misclassifications by the MUAP duration classification model occurred between normal and prolonged MUAPs on the one hand and normal and shortened on the other hand. This misclassification may result from the imperfect reference standard we used for this study or, in fact, the lack of an accepted international reference standard for the evaluation of MUAP characteristics. For example, the proportion of prolonged or shortened MUAPs that may be present in an otherwise healthy muscle is not set, probably because it is very difficult for clinicians to label every MUAP in an n-EMG recording, quantify the proportion of each label and set an optimal cut-off to discern disease from healthy. In our study, we chose to use the label with the highest proportion as an overall label for the entire n-EMG recording while recognizing the limitations of this choice. Fig. 1 indicates that changing this, for example by classifying a muscle abnormal when >25 % of all segments is predicted as having prolonged MUAPs, will have important impact on overall accuracy. There are currently no large n-EMG datasets available that have been labelled by a uniform consensus based reference standard to train future ANN models. As possible other solutions, one or more approaches may be taken. One approach is to train ANN models on quantitative MUAP parameters thereby providing a more objective measure of a specific MUAP characteristics such as duration. Another is to train ANN models on datasets where n-EMG recordings are labeled by a large number of experts and the degree of consensus between experts is taken into account as measure of overall certainty of a label. Finally, ANN models may be trained on datasets where not only MUAP characteristics are used as a reference standard but also the clinical diagnosis, demographics, like age, the specific muscle (including bulbar muscles) and involvement of a muscle in that disorder are used as input (Buchtal and Rosenfalck, 1955). These combined datasets may help to empirically investigate cut-offs for the occurrence of different MUAP characteristics in discerning healthy from disease but may also help to move beyond the currently known and used MUAP characteristics and investigate the diagnostic value of potential other features by using unsupervised learning.

In addition to the limitations concerning the reference standard we employed for the MUAP duration classification model, this study has several other limitations. For the muscle activity classification model, we used a reference standard based on consensus between two investigators which resulted in a reliable but not very efficient reference standard as 25 % of the n-EMG recording was excluded. In the two external datasets, we assumed that neurogenic disorders showed MUAPs of prolonged MUAP duration and myopathic disorders showed MUAPs of shortened duration while

this may be an oversimplification of n-EMG abnormalities that may be found in (some) neurogenic or myopathic disorders. Although this study used the largest dataset so far, we still observed signs of overfitting in model 2 as indicated by some degree of instability during training. As we decided that it was of importance to exclude distant activity, we increased the high-pass filter for the MUAP duration classification model, but this may also have had an impact on MUAP duration of the nearby MUAPs. The acquisition filter settings were largely the same in the datasets used, and we did not investigate potential influences of different low- or high-pass filter settings on our results. We did not investigate signal quality within each of the segments and, for example, excluded distant MUAPs, as we wanted to develop a clinically applicable algorithm. Furthermore, we did not incorporate the presence of abnormal spontaneous activity, number of MUAP phases, MUAP amplitudes, or interference patterns that clinically are used in addition to MUAP duration in the distinction between healthy muscles and muscles affected by neurogenic or myopathic disorders. Finally, we did not directly compare performances of other algorithms with our ANN.

5. Conclusion

We developed a two-step ANN to classify MUAP duration mimicking daily practice in the clinic. The first model discerns rest, muscle contraction and artifacts from real-world n-EMG recordings with very high accuracy. The second model subsequently uses segments predicted as muscle contraction and classifies MUAP duration in these segments as prolonged, shortened or normal with moderate accuracy. Future studies should focus on increasing the number of high quality datasets and solving issues with imperfect reference standards to further build on the potential of AI assisted classification of MUAPs.

Study funding

This study was not funded.

Disclosure

D. Hubers, W. Potters, O. Paalvast, S. de Jonge, B. Doelkahar, L. Wieske, C. Verhamme report no disclosures relevant to the manuscript. M. Tannemaat reports trial support from Argenx and Alexion, consultancies for Argenx and UCB Pharma and research funding from Argenx and NMD Pharma, with all reimbursements received by Leiden University Medical Center.

Acknowledgements

We would like to thank prof. dr. ir. M.J.A.M. van Putten for his input during the design and course of this study. Several authors of this publication are members of the Netherlands Neuromuscular Centre (NL-NMD) and the European Reference Network for rare neuromuscular diseases (ERN-EURO-NMD).

Author contributions

Deborah Hubers: Design or conceptualization of the study; major role in the acquisition of data; analysis or interpretation of the data; drafting or revising the manuscript for intellectual content.

Wouter Potters: Design or conceptualization of the study; major role in the acquisition of data; analysis or interpretation of the data; drafting or revising the manuscript for intellectual content.

Olivier Paalvast: major role in the acquisition of data; analysis or interpretation of the data; drafting or revising the manuscript for intellectual content.

Sterre de Jonge: major role in the acquisition of data; analysis or interpretation of the data; drafting or revising the manuscript for intellectual content.

Brian Doelkahar: analysis or interpretation of the data; drafting or revising the manuscript for intellectual content.

Martijn Tannemaat: major role in the acquisition of data; analysis or interpretation of the data; drafting or revising the manuscript for intellectual content.

Luuk Wieske: Design or conceptualization of the study; major role in the acquisition of data; analysis or interpretation of the data; drafting or revising the manuscript for intellectual content.

Camiel Verhamme: Design or conceptualization of the study; major role in the acquisition of data; analysis or interpretation of the data; drafting or revising the manuscript for intellectual content.

Appendix A. Supplementary material

Supplementary material to this article can be found online at https://doi.org/10.1016/j.clinph.2023.10.008.

References

Bakiya A, Kamalanand K, Rajinikanth V, Nayak RS, Kadry S. Deep neural network assisted diagnosis of time-frequency transformed electromyograms. Multimed Tools Appl 2020;79:11051–67.
Buchtal F, Rosenfalck P. Action potential parameters in different human muscles. Acta Psychiatr Neurol Scand 1955;30:125–31.
Chatterjee S, Roy SS, Bose R, Pratiher S. Feature extraction from multifractal spectrum of electromyograms for diagnosis of neuromuscular disorders. IET Science Meas Technol 2020;14:817–24.
Christodoulou CI, Pattichis CS. Unsupervised pattern recognition for the classification of EMG signals. IEEE Trans Biomed Eng 1999;46:169–78.
Cohen JF, Korevaar DA, Altman DG, Bruns DE, Gatsonis CA, Hooft L, et al. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. BMJ Open 2016;6:e012799.
Doulah ABMSU, Fattah SA. Neuromuscular disease classification based on mel frequency cepstrum of motor unit action potential. In: 2014 Int Conf Electr Eng Inf Commun Technol. p. 1–4.
Farkas C, Hamilton-Wright A, Parsaei H, Stashuk DW. A review of clinical quantitative electromyography. Crit Rev Biomed Eng 2010;38:467–85.
Fuglsang-Frederiksen A. The role of different EMG methods in evaluating myopathy. Clin Neurophysiol 2006;117:1173–89.
Katirji B. In: Neuromuscular disorders in clinical practice. New York: Springer; 2014. p. 3–21.
Katsis CD, Exarchos TP, Papaloukas C, Goletsis Y, Fotiadis DI, Saras I. A two-stage method for MUAP classification based on EMG decomposition. Comput Biol Med 2007;37:1232–40.
Kendall R, Werner RA. Interrater reliability of the needle examination in lumbosacral radiculopathy. Muscle Nerve 2006;34:238–41.
Meinsma G, Heida C, van Putten MJAM. Advanced techniques for signal analysis. Enschede: University of Twente; 2019.
Mishra VK, Bajaj V, Kumar A, Singh GK. Analysis of ALS and normal EMG signals based on empirical mode decomposition. IET Sci Meas Technol 2016;10:963–71.
Mokdad A, Debbal SMEA, Meziani F. Diagnosis of amyotrophic lateral sclerosis (ALS) disorders based on electromyogram (EMG) signal analysis and feature selection. Polish J Med Phys Eng 2020;26:155–60.
Nagineni S, Taran S, Bajaj V. Features based on variational mode decomposition for identification of neuromuscular disorder using EMG signals. Health Inf Sci Syst 2018;6:1–10.
Nikolic M. Detailed analysis of clinical electromyography signals EMG decomposition, findings and firing pattern analysis in controls and patients with myopathy and amytrophic lateral sclerosis PhD Thesis. Faculty of Health Science, University of Copenhagen; 2001.
Nodera H, Osaki Y, Yamazaki H, Mori A, Izumi Y, Kaji R. Deep learning for waveform identification of resting needle electromyography signals. Clin Neurophysiol 2019;130:617–23.
Preston DC, Shapiro BE. Electromyography and neuromuscular disorders: clinical electrophysical correlations. In: Basic electromyography: Analysis of motor unit action potentials. Philadelphia: Elsevier Saunders; 2013. p. 235–49.
Rao KS, Manjunath K. Speech recognition using articulatory and excitation source features. Cham: Springer International Publishing AG; 2017.

Sahidullah M, Saha G. Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition. Speech Comm 2012;54:543–65.
Samanta K, Roy SS, Modak S, Chatterjee S, Bose R. Neuromuscular disease detection employing deep feature extraction from cross spectrum images of electromyography signals. In: Annu int conf IEEE eng med biol soc. p. 694–7.
Sengur A, Akbulut Y, Guo Y, Bajaj V. Classification of amyotrophic lateral sclerosis disease based on convolutional neural network and reinforcement sample learning algorithm. Health Inf Sci Syst 2017;5:1–7.
Subasi A. Classification of EMG signals using combined features and soft computing techniques. Appl Soft Comput 2012;12:2188–98.
Szegedy C, Loffe S, Vanhoucke V, Alemi AA. Inception-v4, inception-resnet and the impact of residual connections on learning. In: Proceedings of the thirty-first AAAI conference on artificial intelligence. p. 4278–84.
Tankisi H, Burke D, Cui L, Carvalho M, Kuwabara S, Nandedkar SD, et al. Standards of instrumentation of EMG. Clin Neurophysiol 2020;131:243–58.
Tannemaat MR, Kefalas M, Geraedts VJ, Remijn-Nelissen L, Verschuuren AJM, Koch M, et al. Distinguishing normal, neuropathic and myopathic EMG with an automated machine learning approach. Clin Neurophysiol 2023;146:49–54.
Umesh S, Cohen L, Nelson DJ. Fitting the Mel scale. 1999 IEEE Int Conf Acoust Speech Signal Process Proc. ICASSP99 (Cat. No. 99CH36258) 1999;1. p. 217–20.
Zapf A, Castell S, Morawietz L, Karch A. Measuring inter-rater reliability for nominal data–which coefficients and confidence intervals are appropriate? BMC Med Res Method 2016;16:1–10.