Distinguishing normal, neuropathic and myopathic EMG with an automated machine learning approach

Tannemaat, M.R.; Kefalas, M.; Geraedts, V.J.; Remijn-Nelissen, L.; Verschuuren, A.J.M.; Koch, M.; ... ; Bäck, T.H.W.

Clinical Neurophysiology 146 (2023) 49–54

Contents lists available at ScienceDirect

Clinical Neurophysiology

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

Distinguishing normal, neuropathic and myopathic EMG with an automated machine learning approach

M.R. Tannemaat [a,*], M. Kefalas [b], V.J. Geraedts [a,c], L. Remijn-Nelissen [a], A.J.M. Verschuuren [a], M. Koch [b], A.V. Kononova [b], H. Wang [b], T.H.W. Bäck [b]

[a] Leiden University Medical Centre, Department of Neurology, The Netherlands
[b] Leiden Institute of Advanced Computer Science, The Netherlands
[c] Leiden University Medical Centre, Department of Clinical Epidemiology, The Netherlands

HIGHLIGHTS

- A machine learning algorithm can differentiate EMGs from healthy individuals from patients with ALS with a high diagnostic yield.
- The automated approach aimed at limiting all arbitrary choices with regards to epoch selection and hyperparameter optimization.
- This algorithm allows the identification of features used for classification, allowing interpretation of the model.

ARTICLE INFO

Article history:
Accepted 26 November 2022
Available online 9 December 2022

Keywords:
Machine learning
Quantitative EMG
Amyotrophic lateral sclerosis
Inclusion body myositis

ABSTRACT

Objective: Distinguishing normal, neuropathic and myopathic electromyography (EMG) traces can be challenging. We aimed to create an automated time series classification algorithm.
Methods: EMGs of healthy controls (HC, n = 25), patients with amyotrophic lateral sclerosis (ALS, n = 20) and inclusion body myositis (IBM, n = 20), were retrospectively selected based on longitudinal clinical follow-up data (ALS and HC) or muscle biopsy (IBM). A machine learning pipeline was applied based on 5-second EMG fragments of each muscle. Diagnostic yield expressed as area under the curve (AUC) of a receiver-operator characteristics curve, accuracy, sensitivity, and specificity were determined per muscle (muscle-level) and per patient (patient-level).
Results: Diagnostic yield of the classification ALS vs. HC was: AUC 0.834 ± 0.014 at muscle-level and 0.856 ± 0.009 at patient-level. For the classification HC vs. IBM, AUC was 0.744 ± 0.043 at muscle-level and 0.735 ± 0.029 at patient-level. For the classification ALS vs. IBM, AUC was 0.569 ± 0.024 at muscle-level and 0.689 ± 0.035 at patient-level.
Conclusions: An automated time series classification algorithm can distinguish EMGs from healthy individuals from those of patients with ALS with a high diagnostic yield. Using longer EMG fragments with different levels of muscle activation may improve performance.
Significance: In the future, machine learning algorithms may help improve the diagnostic accuracy of EMG examinations.

1. Introduction

Needle electromyography (EMG) is a technique in which an electrode is inserted into the muscle to record and evaluate the electrical activity of the muscle at rest and during voluntary contraction. Interpretation is usually based on qualitative visual assessment of the signal (Dumitru et al., 2001). This approach is subjective, laborious, and its reliability depends substantially on the examiner's experience. Moreover, even among experts, inter-rater agreement is poor (Narayanaswami et al., 2016). The diagnostic yield of quantitative EMG (qEMG) methods such as turns-amplitude analysis, is similar to visual inspection (Daube and Rubin, 2009; Thornton and Michell, 2012) and the limitations of EMG equipment currently preclude the use of more advanced techniques for routine examinations (Thornton and Michell, 2012).

A neuropathic EMG, with high-amplitude and long duration motor unit action potentials (MUAPs), and a reduced interference pattern should theoretically be clearly distinguishable from a myopathic EMG containing smaller, short-duration polyphasic MUPs

* Corresponding author.
E-mail address: m.r.tannemaat@lumc.nl (M.R. Tannemaat).

https://doi.org/10.1016/j.clinph.2022.11.019
1388-2457/

and a full interference pattern. In practice, however, differentiation between a neuropathy and a myopathy can be challenging. This challenge is exemplified by the difficulties in distinguishing EMGs obtained from patients with amyotrophic lateral sclerosis (ALS) from EMGs obtained from those with inclusion body myositis (IBM) (Badrising et al., 2005). Both diseases are characterized clinically by progressive weakness without sensory deficits, but whereas ALS is a progressive neurodegenerative disease and fatal within several years, life expectancy is not affected in IBM (Badrising et al., 2005).

Recent advances in computer processing power and machine learning (ML) techniques enable processing a large number of features without any underlying assumptions about the nature of the signal. We have previously shown that such an approach, developed for the automotive industry but applied to electroencephalography (EEG) signals, could classify Parkinson's disease patients with good cognition from those with poor cognition with an accuracy of 91% (Geraedts et al., 2021a). Here, we aimed to evaluate an automated time series classification algorithm to differentiate EMG time series from healthy individuals, patients with a neuropathic disease (ALS) or a myopathy (IBM). With this approach, we aimed to limit all arbitrary choices with regards to hyperparameter optimization.

2. Methods

2.1. EMG acquisition and storage

All EMGs were recorded at the department of clinical neurophysiology of Leiden University Medical Center (LUMC), a tertiary referral center for neuromuscular diseases, during the period 2004–2019. All EMGs were performed with concentric needle electrodes with a low-frequency (high pass) filter of 30 Hz and a high-frequency (low-pass) filter of 3 kHz and recorded using Medelec Synergy electromyography equipment (Oxford Instruments, Abingdon, Oxfordshire, UK). This equipment routinely stores the last 40 seconds of an EMG. In general, the assessment takes place in three phases: with the muscle at rest, during slight activation and during (near-) maximal activation. Recording at maximal muscle activation is commonly avoided when the EMG signal appears to be normal at near-maximal activation levels, as the EMG becomes increasingly painful when the muscle is fully activated. All data were gathered in the process of routine clinical care and a formal ethical approval was therefore waived for this particular study. Dutch law does not require individual informed consent for the use of anonymized data for retrospective scientific research.

2.2. Selection of patients

Patient selection was retrospective and constituted a sample of convenience, based on the availability of data. This was unavoidable due to the rarity of IBM and our stringent use of selection criteria. Sample size for the ALS and control group were matched to the IBM sample size. Selection criteria per group were: For IBM, all patients diagnosed with IBM were identified from searched electronic patient records for the period 2004–2019. We then selected all patients with muscle biopsies showing atrophy, inflammation, and rimmed vacuoles (Hilton-Jones and Brady, 2016) and from whom EMG data were available. For ALS, we used electronic patient records to select patients with definite ALS following the Awaji criteria, based on typical clinical findings, confirmed with EMG (Costa et al., 2012) (although not necessarily on the first EMG, which was included in the analysis described here). In all cases, correctness of the diagnosis was confirmed by a clinical course of neurological deterioration leading to death. Healthy control subjects (HC) were defined as subjects who fulfilled all of the following criteria: (1) the subject was considered unlikely to suffer from a neurological disease by the treating neurologist, presenting with non-specific complaints (e.g., muscle aches, pain, or fear of a neuromuscular disease), (2) no muscle weakness upon neurological examination, (3) no abnormalities suggesting a neuromuscular disease in any available ancillary investigation, and (4) no signs of muscle weakness during a follow-up period of at least two years.

For a comparison of the accuracy of the ML pipeline with the results of the clinical conclusion at the time of the evaluation, we extracted the conclusions from all EMG reports and classified them in three groups: (1) unambiguously correct (e.g., "no abnormalities" for a subject from the HC group, or "myopathic changes" for a subject in the IBM group), (2) unambiguously incorrect (e.g., "neurogenic changes" for a subject in the IBM group and (3) ambiguous conclusions (e.g., "findings suggesting a loss of motorneurons, consider repeating the examination at a later time" for a subject in the ALS group). We then calculated the accuracy of the clinical report twice: once with the ambiguous results included in the group of correct diagnoses, and once with the ambiguous results included in the group of incorrect diagnoses, to provide the upper and lower limits of the accuracy of the clinical report.

2.3. Selection of EMG traces

When multiple EMG examinations were available from one patient, we selected the first (oldest). We assumed that the earliest EMG in the disease course is both the most challenging, as abnormalities are likely to become more pronounced as the disease progresses and the most important in terms of clinical impact and prognostication. From the selected examination, we extracted EMG traces from all investigated muscles. As recordings were made in the process of routine clinical care, it is likely that the selection of muscles for investigation depended on the clinical query and differed between groups (i.e., bulbar muscles were probably more likely to be included in the examination in cases with a high suspicion of ALS). A complete list of investigated muscles per disease group is available in Supplemental Table 1. From each muscle, we visually inspected the raw trace and selected the longest available, continuous fragment containing MUPs without needle movement, 50 Hz artefacts from wall outlets or other artefacts. From each EMG fragment, we clipped the last 5 seconds and used this for automated analysis. In general, these fragments were likely to contain EMG activity at near-maximal activation but selection based on activation level did not take place. Some traces were recorded with a sample frequency of 5000 Hz, others with a frequency of 4800 Hz. Sample frequency was not related to disease class. All data were downsampled to 4800 Hz prior to further processing.

2.4. Machine learning pipeline

A previously reported ML pipeline approach was used for time series classification purposes (Geraedts et al., 2021a, 2021b; Kefalas et al., 2020). The ML pipeline consists of four phases: (1) feature-extraction, (2) feature selection, (3) training of a classifier, and (4) hyperparameter optimization. All four steps are completely automated, with the EMG time series as the sole input and the class-labels (i.e. control, ALS patient or IBM) as output. No information on the level of activation or results of visual qualitative inspection of the EMG were used as input data. The library 'Time Series FeatuRe Extraction on basis of Scalable Hypothesis tests' (TsFresh) was used to extract features from the time series (Christ et al., 2018), resulting in 794 features per EMG fragment. In this context, features are best understood as properties of a time series which may contain information on the nature of the signal. The library TsFresh contains a large array of potentially relevant features, including basic features such as the mean, maximum or variance

of a dataset, but also more compel aspects including Fourier analyses, entropy and skewness.

Feature selection was performed using the Boruta algorithm, by testing the variable importance (VIMP) of each feature against that of 'shadow features', which are created by random shuffling of the original features. The VIMP of shadow and original are obtained from a random forest model trained thereon. An original feature would be selected if its VIMP frequently dominated the maximal VIMP of shadow features, in multiple independent trials (Kursa and Rudnicki, 2010). After feature-selection, the feature set is used to train a Random Forest Classifier (RFC). An RFC is an ensemble of decision trees; the resulting decision is the majority vote from all decision trees (Friedman et al., 2001). To ensure generalizability of the RFC, a cross-validation procedure was adopted: the data were randomly split into ten folds, after which training was performed on nine folds and tested on the remaining fold. This process was repeated until each fold had served as a test set; the average of all test scores of the computations represented the final score. To quantify the reliability of the end result, this process was repeated five times and the results were averaged again. The RFC was executed 100 times per fold and the result was averaged. The hyperparameters of the RFC, such as the number of decision trees and their individual tree depths, are optimized with a variant of the Bayesian Optimization technique called Mixed Integer Parallel Efficient Global Optimization (MIP-EGO) for mixed-integer categorical search spaces (Wang et al., 2018; Wang et al., 2017). To optimize the hyperparameters of the RFC, MIP-EGO optimized the F1-macro score of another (nested) 10-fold cross-validation. This nested cross-validation was executed on the training-fold of each split of the overall 10-fold cross-validation process. The F1-macro score is defined as the average of the F1 score of each class (Kefalas et al., 2020). We used the F1-score to find the balanced score between the two classes being considered (ALS vs. HC, ALS vs. IBM, IBM vs. HC). both cross-validations were stratified in order to preserve the percentage of samples for each class.

2.5. Muscle-level and patient-level approach.

As described previously (Kefalas et al., 2020), the pipeline was trained by using the diagnosis of each patient as classification gold standard for all EMG traces belonging to that patient, resulting in a probability score for each trace. Diagnostic yield was subsequently evaluated twice: for the muscle-level approach, we treated each trace as an independent observation for receiver-operator characteristics (ROC) analysis to calculate accuracy, sensitivity, and specificity, based on probability scores. For the patient-level approach, we calculated the median of the prediction probabilities of all muscles from the same patient to make a patient-level predictive decision.

2.6. Feature importance

For each comparison (HC vs. ALS, HC vs. IBM, and ALS vs. IBM), we created a list of features that appeared in the final decision tree in each of the 10-fold cross-validations. We repeated this by selecting the features in each of the 5 repetitions to create a final list of all features that were retained in each comparison, assuming that these were the most relevant.

2.7. Statistical analysis

Demographic and clinical variables were compared between the HC, ALS and IBM groups using one-way ANOVA if normally distributed and Kruskal-Wallis tests if not normally-distributed in continuous variables and Pearson's χ^2 tests in case of categorical data.

All statistical analyses not pertaining to the automated ML pipeline (i.e. baseline characteristics) were performed using IBM Statistical Package for the Social Sciences (SPSS) 25 Software (SPSS inc. Chicago, Illinois, USA).

3. Results

3.1. Patient characteristics

Baseline characteristics for all three groups are provided in Table 1. A total of 65 patients were included, 20 patients with ALS, 20 patients with IBM, and 25 healthy controls. Healthy controls were significantly younger than patients with ALS and IBM (p =< 0.001). A total number of 380 muscles were included in the measurements. More muscles were measured in the ALS group than in HC and IBM groups. The accuracy of the clinical EMG report ranged from 96-100% for the HC group, 65–80% in the IBM group and 45–85% in the ALS group.

3.2. Diagnostic yield of the ML pipeline

The performance scores for both the muscle-level as the patient-level approach are shown in Table 2. The muscle-level approach treats each muscle as a separate observation; for the patient-level approach, the median probability score of all examined muscles for that patient was calculated. Each resulting performance score represents the average of 5 independent runs of the automated machine learning pipeline. For all three classification tasks, accuracy and diagnostic yield (AUC of the ROC curve) were highly similar across all five repetitions and for both the muscle-level and patient-level approach (Fig. 1). The highest diagnostic yield was reached for the classification ALS vs. HC: AUC was 0.83 4 ± 0.014 at the muscle level and 0.856 ± 0.009 at the patient level. The lowest diagnostic yield was reached for the classification ALS vs. IBM: AUC was 0.569 ± 0.024 at the muscle level and 0.689 ± 0. 035 at the patient level.

3.3. Feature importance

For the classification ALS vs. HC, 12 significant features were retained in all cross-validations and repetitions. These features, derived from the online feature library "TsFresh" (Christ et al., 2018) are shown in Fig. 2. Detailed descriptions can be found online, but in brief, four features were related to the presence of reoccurring values in the data series: (1) the percentage of values that were present in the time series more than once", (2) a factor which is 1 if all values in the time series occur only once, and below one if this is not the case, (3) the sum of all data points that are present in the time series more than once and (4) percentage of non-unique data points. Four features were related to spectral measures of the signal: (1) the absolute value of the 1st coefficient of the 1D discrete Fourier transformation, (2) the real part of the 1st coefficient of the 1D discrete Fourier transformation, (3) the absolute value of the 36rd coefficient of the 1D discrete Fourier transformation and (4) cross power spectral density of the time series x at different frequencies (the time series is first shifted from the time domain to the frequency domain. Three features were related to the amplitude of the signal: (1) the sum of all time series values, (2) the standard deviation of all values and (3) the absolute energy of the time series, i.e. the sum over the squared values. One and one feature represented the frequency of value "0". For the classification IBM vs. HC, two relevant features were retained: "number_crossing_m__m_0", which represents the number of zero crossings in a signal, with a feature importance of 4.6 ± 0.1 and "fft_coeffi cient__coeff_34__attr_"abs"", which represents the absolute value

Table 1
Demographics and clinical characteristics.

	HC (*n* = 25)	IBM (*n* = 20)	ALS (*n* = 20)	*p* value
Age in years, mean ± SD[a]	51 ± 12	70 ± 7	64 ± 6	<0.001
Female sex, *n* (%)[b]	11 (44)	9 (45)	9 (45)	0.997
Symptom duration, median (IQR)[c]	2 (0.5–10)	3 (2–10)	2 (1–2.75)	0.109
muscles investigated, median (IQR)[c]	4 (3–7)	5 (3–8)	7 (6–10.5)	0.002
Original report correct, *n* (%)[d]	24–25 (96–100)	13–16 (65–80)	9–17 (45–85)	

[a] One-way ANOVA; [b] Pearson's χ^2 test; [c] Kruskal-Wallis test. SD: standard deviation, IQR: interquartile range. HC: healthy controls, IBM: inclusion body myositis, ALS: amyotrophic lateral sclerosis.
There were significant differences in patient age and number of muscles investigated between groups. [d] Range of clinical EMG report: lower bound calculated by interpreting ambiguous test results as incorrect; upper bound calculated by interpreting ambiguous results as correct.

Table 2
Machine learning model performance scores.

	ALS vs. HC		IBM vs. HC		ALS vs. IBM	
	Muscle-level	**Patient-level**	**Muscle-level**	**Patient-level**	**Muscle-level**	**Patient-level**
Accuracy	0.779 ± 0.019	0.779 ± 0.025	0.671 ± 0.023	0.684 ± 0.029	0.547 ± 0.170	0.570 ± 0.048
Sensitivity	0.842 ± 0.025	0.850 ± 0.035	0.664 ± 0.030	0.640 ± 0.022	0.685 ± 0.019	0.860 ± 0.055
Specificity	0.688 ± 0.038	0.704 ± 0.036	0.672 ± 0.027	0.720 ± 0.040	0.359 ± 0.020	0.280 ± 0.144
AUC	0.834 ± 0.014	0.856 ± 0.009	0.744 ± 0.043	0.735 ± 0.029	0.569 ± 0.024	0.689 ± 0.035

AUC: Area under the curve for the receiver operator characteristics curve. HC: healthy controls, IBM: inclusion body myositis, ALS: amyotrophic lateral sclerosis. Values indicate mean +/- standard deviation for 5 repetitions of the pipeline, each consisting of 10 cross-validations. Muscle-level and patient-level analysis yielded highly similar values.

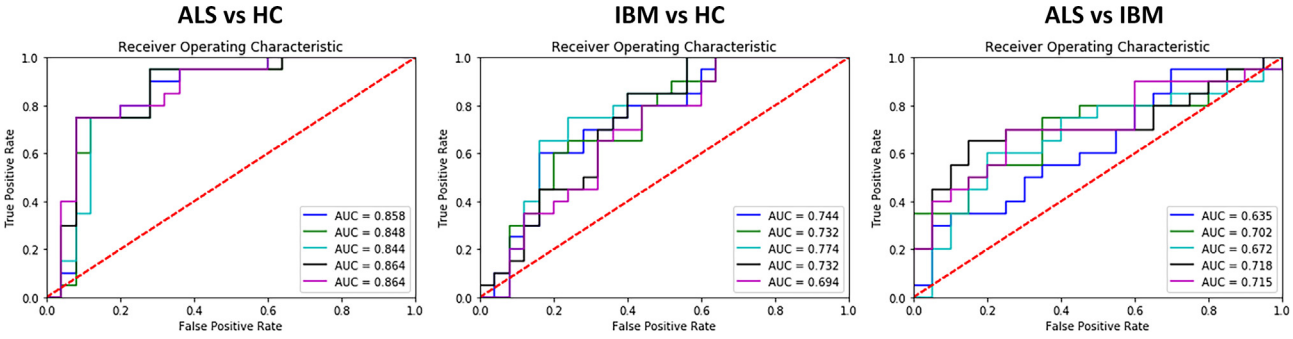

Fig. 1. ROC curves at the patient-level. ROC: receiver operator characteristics. HC: healthy controls, IBM: inclusion body myositis, ALS: amyotrophic lateral sclerosis. AUC: Area under the curve. Each colored line represents one of five repetitions of a 10-fold cross-validation. Mean AUC values are provided in Table 2. The highest diagnostic yield was reached for the comparison ALS vs. HC.

of the 35th coefficient of the 1D discrete Fourier Transformation (importance 4.0 ± 0.1). For the classification ALS vs. IBM, only one feature was consistently present in all 5 repetitions: "ar_coeffi cient__k_10__coeff_3", which represents the third coefficient of the lag 10 autoregression process, with a mean importance score of 15.8 ± 2.1.

4. Discussion

In this study, we show that an automated time series classification algorithm can differentiate EMGs from healthy individuals from those of patients with ALS with a high diagnostic yield. For the diagnosis ALS, the accuracy of the ML pipeline and the clinical EMG reports at the time of the investigation appeared to be within the same range, although a statistical analysis was not possible due to the qualitative nature of the latter. Diagnostic yield for the distinction between IBM and healthy controls, and between ALS and IBM was somewhat lower. This is in line with clinical experience: neurogenic abnormalities are relatively easy to distinguish from a normal EMG. The lowest diagnostic yield was reached for the distinction between ALS and IBM. This is also similar to clinical practice, as the distinction between these diseases on EMG data can be notoriously difficult. A retrospective study of mislabeled IBM patients found that routine EMG commonly pointed to a neurogenic disorder: it showed fibrillation potentials and positive sharp waves, as well as fasciculation potentials and excessive amounts of polyphasic long-duration "neurogenic" MUAPs in the majority of mislabeled patients (Dabby et al., 2001). ALS and IBM were chosen here as exemplary diseases for neuropathy and myopathy, because they can be diagnosed based on clearly defined criteria, thus minimizing diagnostic uncertainty. Our assumption was that an algorithm capable of distinguishing ALS from IBM would be useful for the identification of other neurogenic changes (e.g., traumatic nerve injury, radiculopathy, polyneuropathy) and myopathies as well. As the ultimate aim was to develop an ML approach to distinguish neuropathic and myopathic EMG signals, we did not include clinical data in the analysis, other than to confirm the diagnosis.

Previously reported machine learning algorithms differentiating healthy subjects from patients with myopathy and neuropathy reported accuracies between 86.3% and 99% (Artuğ et al., 2014;

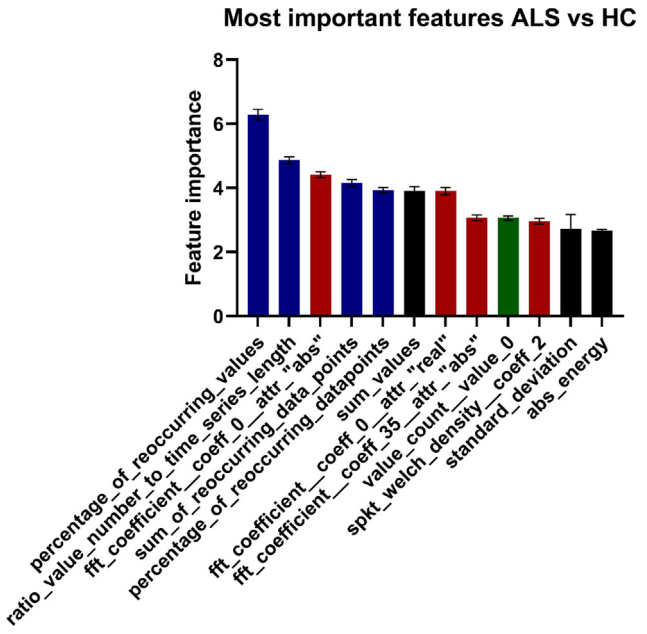

Fig. 2. Mean feature importance of consistently retained features for the classification between ALS and HC. HC: healthy controls, ALS: amyotrophic lateral sclerosis. Features are derived from TsFresh, an online feature library (Christ et al., 2018). Detailed descriptions can be found online, but in brief, the meaning of all features from left to right is "percentage of values that are present in the time series more than once", "factor which is 1 if all values in the time series occur only once, and below one if this is not the case", "absolute value of the 1st coefficient of the 1D discrete Fourier Transformation", "sum of all data points, that are present in the time series more than once", "percentage of non-unique data points.", "sum of all time series values", "real part of the 1st coefficient of the 1D discrete Fourier Transformation", "absolute value of the 36rd coefficient of the 1D discrete Fourier Transformation", "count occurrences of value 0", "cross power spectral density of the time series x at different frequencies (the time series is first shifted from the time domain to the frequency domain)", "standard deviation of the signal", "absolute energy of the time series, i.e. the sum over the squared values". Blue bars indicate features related to the presence of reoccurring values in the data series, red bars are related to spectral measures, black bars represent features with the amplitude of the signal and the green bar represents the remaining feature, which indicates the frequency of value "0" in the signal.

Dobrowolski et al., 2012; Elamvazuthi et al., 2015; Istenic et al., 2010; Mokdad et al., 2020; Naik et al., 2016; Sengur et al., 2017; Subasi, 2012, 2018). However, these studies lacked a clear description on the criteria used to diagnose IBM, ALS, and healthy subjects, and generally involved small numbers of patients or simulated data.

The main strength of our study is the stringent use of clinical criteria and gold standards that did not reply on subjective expert opinion. In addition, classification was completely independent of the EMG signal used for the machine learning pipeline: for IBM, the gold standard was muscle biopsy and for ALS and healthy controls, longitudinal clinical follow-up. Furthermore we used the earliest available EMG recording. This is the most clinically relevant recording, but it is likely more difficult to differentiate EMGs from ALS or IBM patients at early disease stages from healthy subjects. Diagnostic yield would probably have been better if EMGs from more advanced disease stages would have been used, although this would have been less relevant for clinical practice. Another strength of this study is the use of a sophisticated ML pipeline with automated hyperparameter optimization, fitting our aim to limit the use of arbitrary choices as much as possible. In line with this aim, we did not specifically select EMG fragments based on the level of muscle activation, as this would require subjective interpretation. Selection of longer traces, containing the EMG signal from muscles at different stages of activation would likely have improved diagnostic yield further, especially if these traces could be labeled during the examination to indicate the trace contained a recording of the muscle at rest, at minimal contraction or at (near) maximal contraction.

In comparison to deep learning algorithms, a major advantage of the ML approach described here is its ability to identify the relative importance of features used for classification, allowing interpretation of the model. In particular for the classification "ALS vs. HC", a number of features were retained in all cross-validations and all repetitions. Although these data should be interpreted with caution due to the relatively small number of included subjects, there are some suggestions that the most important features were indeed indicative of abnormalities that are commonly used in the clinical interpretation of the EMG. Four features were related to the percentage of reoccurring values in the signal, which may have been caused by reduced recruitment patterns consisting of a small number of repeating motor unit potentials. Four features related to spectral measures may be influenced by MUAP duration. Three features related to power or amplitude of the signal may have been related to increased MUAP amplitudes observed in ALS, and the final feature, which represented the frequency of value zero in the traces, may be affected by the reduced interference pattern that is commonly seen in ALS. For future research, a more in-depth analysis of which features are relevant for disease classification could lead to the discovery of novel biomarkers for disease classification and progression.

It is remarkable that performance at the individual muscle-level (in an approach that treated each muscle as a separate observation) was similar to performance at the patient-level (in which all muscles from each patient were combined). Given the "patchy" nature of ALS, especially at early disease stages, some EMG traces from ALS patients were likely to be from apparently unaffected muscles and thus would appear to resemble normal EMG traces. Therefore, prior to this analysis we expected the patient-level approach to reach a higher diagnostic yield. It should be noted that in both the ALS and IBM groups, all muscles were included in both the training and the validation set, regardless of whether they appeared to be affected. An uneven distribution of the level of disease activity would therefore affect both the muscle and the patient level analysis, as both are based on the same ML algorithm; the patient level approach is essentially a post-processing step to calculate the average score of all muscles from each patient. To investigate whether the algorithm would improve when only clearly abnormal traces were included, we performed an additional analysis using only EMG traces which were labelled neuropathic or myopathic in the original clinical report. Unfortunately, this analysis did not yield meaningful results (data not shown), probably because this model was trained on a much smaller and imbalanced dataset and relied on a subjective assessment.

Limitations of this study are the small data set, which is inevitable given the rarity of IBM in particular, and the relatively short fragments of EMG recordings used. Selection of the fragment was based on the absence of artifacts, although we aimed to select the last 5 seconds of every EMG signal assuming that this segment, usually containing (near-) maximal muscle contraction, would contain the most useful information for classification. The ML algorithm was validated through a 10-fold cross-validation. Although this approach reduces the risk of overfitting, a limited degree of overfitting cannot be ruled out completely. Due to the small sample size used here, the differences in the number and type of examined muscles between different disease classes and the (albeit small) risk of overfitting, the data presented here are best considered a 'proof of concept', showing the potential of an automated approach. However, prior to clinical application, further refinement of the algorithm and validation on external data, preferably from another hospital, would be needed to assess generalizability.

Ultimately, our findings could lead to the development of an ML pipeline integrated into EMG software for automatic classification of the signal. Such a tool could improve diagnostic accuracy, reduce inter-observer variability and reduce the need for assessment by a trained clinical neurophysiologist, which is laborious, expensive, and time-consuming.

Data availability statement

The data that support the findings of this study are available from the corresponding author, upon reasonable request.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Ethics approval statement

Formal approval waived by ethics board due to retrospective observational nature of the study.

Patient consent statement

Not applicable.

Permission to reproduce material from other sources

Not applicable.

Clinical Trial Registration

Not applicable.

Conflict of Interest

None of the authors report any potential conflict of interest.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.clinph.2022.11.019.

References

Artuğ T, Goker I, Bolat B, Tulum G, Osman O, Baslo M. Feature extraction and classification of neuromuscular diseases using scanning EMG. 2014 IEEE International Symposium on Innovations in Intelligent Systems and Applications (INISTA) Proceedings; 2014. p. 262–5.

Badrising UA, Maat-Schieman ML, van Houwelingen JC, van Doorn PA, van Duinen SG, van Engelen BG, et al. Inclusion body myositis. Clinical features and clinical course of the disease in 64 patients. J Neurol 2005;252(12):1448–54.

Christ M, Braun N, Neuffer J, Kempa-Liehr A. Time Series FeatuRe Extraction on basis of Scalable Hypothesis tests (tsfresh – A Python package). Neurocomputing 2018;307:72–7.

Costa J, Swash M, de Carvalho M. Awaji criteria for the diagnosis of amyotrophic lateral sclerosis:a systematic review. Arch Neurol 2012;69(11):1410–6.

Dabby R, Lange DJ, Trojaborg W, Hays AP, Lovelace RE, Brannagan TH, et al. Inclusion body myositis mimicking motor neuron disease. Arch Neurol 2001;58(8):1253–6.

Daube JR, Rubin DI. Needle electromyography. Muscle Nerve 2009;39(2):244–70.

Dobrowolski AP, Wierzbowski M, Tomczykiewicz K. Multiresolution MUAPs decomposition and SVM-based analysis in the classification of neuromuscular disorders. Comput Methods Programs Biomed 2012;107(3):393–403.

Dumitru D, Amato A, Zwarts MJ. Electrodiagnostic Medicine. 2nd ed. Philadelphia: Hanley & Belfus; 2001.

Elamvazuthi I, Duy NHX, Ali Z, Su SW, Khan MKAA, Parasuraman S. Electromyography (EMG) based Classification of Neuromuscular Disorders using Multi-Layer Perceptron. Proc Comput Sci 2015;76:223–8.

Friedman J, Hastie T, Tibshirani R. The elements of statistical learning: Springer Series in Statistics. New York: Springer; 2001.

Geraedts VJ, Koch M, Contarino MF, Middelkoop HAM, Wang H, van Hilten JJ, et al. Machine learning for automated EEG-based biomarkers of cognitive impairment during Deep Brain Stimulation screening in patients with Parkinson's Disease. Clin Neurophysiol 2021a;132(5):1041–8.

Geraedts VJ, Koch M, Kuiper R, Kefalas M, Bäck THW, van Hilten JJ, et al. Preoperative Electroencephalography-Based Machine Learning Predicts Cognitive Deterioration After Subthalamic Deep Brain Stimulation. Mov Disord 2021b;36(10):2324–34.

Hilton-Jones D, Brady S. Diagnostic criteria for inclusion body myositis. J Int Med 2016;280(1):52–62.

Istenic R, Kaplanis PA, Pattichis CS, Zazula D. Multiscale entropy-based approach to automated surface EMG classification of neuromuscular disorders. Med Biol Eng Compu 2010;48(8):773–81.

Kefalas M, Koch M, Geraedts VJ, Wang H, Tannemaat MR, Bäck THW. Automated Machine Learning for the Classification of Normal and Abnormal Electromyography Data. 2020 IEEE International Conference on Big Data (Big Data); 2020. p. 1176–85.

Kursa M, Rudnicki W. Feature Selection with Boruta Package. J Stat Softw 2010;36:1–13.

Mokdad A, Debbal SMEA, Meziani F. Diagnosis of amyotrophic lateral sclerosis (ALS) disorders based on electromyogram (EMG) signal analysis and feature selection. Polish J Medical Phys Eng 2020;26(3):155–60.

Naik GR, Selvan SE, Nguyen HT. Single-Channel EMG Classification With Ensemble-Empirical-Mode-Decomposition-Based ICA for Diagnosing Neuromuscular Disorders. IEEE Trans Neural Syst Rehabil Eng 2016;24(7):734–43.

Narayanaswami P, Geisbush T, Jones L, Weiss M, Mozaffar T, Gronseth G, et al. Critically re-evaluating a common technique: Accuracy, reliability, and confirmation bias of EMG. Neurology 2016;86(3):218–23.

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):9.

Subasi A. Medical decision support system for diagnosis of neuromuscular disorders using DWT and fuzzy support vector machines. Comput Biol Med 2012;42(8):806–15.

Subasi A, Yaman E, Somaily Y, Alynabawi H, Alobaidi F, Altheibani S. Automated EMG Signal Classification for Diagnosis of Neuromuscular Disorders Using DWT and Bagging. Proc Comput Sci 2018;140:230–7.

Thornton RC, Michell AW. Techniques and applications of EMG: measuring motor units from structure to function. J Neurol 2012;259(3):585–94.

Wang H, Emmerich M, Bäck T. Cooling Strategies for the Moment-Generating Function in Bayesian Global Optimization. 2018 IEEE Congress on Evolutionary Computation (CEC); 2018. p. 1–8. https://ieeexplore.ieee.org/document/8477956.

Wang H, van Stein B, Emmerich M, Bäck T. A new acquisition function for Bayesian optimization based on the moment-generating function. IEEE International Conference on Systems, Man, and Cybernetics (SMC); 2017. p. 507–12.