Interpretable machine learning for the diagnosis of hyperkinetic movement disorders

Elina L. van den Brandhof[1,2], Jan W.J. Elting[2], Inge Tuitert[2], Jelle R. Dalenberg[2],
A.M. Madelein van der Stouwe[2], Marina A.J. Tijssen[2] and Michael Biehl[1,3]

1 - Univ. of Groningen - Bernoulli Institute for Mathematics,
Computer Science and Artificial Intelligence, Groningen, NL

2 - Univ. Medical Center Groningen (UMCG), (a) Dept. of Neurology,
(b) Expertise Center Movement Disorders Groningen, Groningen, NL

3 - Dept. of Metabolism and Systems Science, College of Medicine
and Health, Univ. of Birmingham, UK

Abstract. We present a machine learning approach to the challenging differentiation of hyperkinetic movement disorders, based on accelerometric sensor data. We address the diagnosis of essential tremor and cortical myoclonus as a specific example. Generalized Matrix Relevance Learning Vector Quantization (GMLVQ) systems are applied directly to power spectra obtained from eight sensors recording upper body movements. We find excellent validation performance of the classifiers. Moreover, GMLVQ provides insight into the characteristic patterns of the phenotypes and the importance of particular frequency ranges in the spectra. We demonstrate that the explanatory power of the classifier is further enhanced when integrating information from several tasks per subject.

1 Introduction

Hyperkinetic movement disorders cause involuntary movements that can affect various parts of the body. As important examples, tremor is characterized by oscillatory movements around joints [1], while myoclonus causes sudden shock-like jerks resulting from muscular contractions [2]. Both conditions predominantly affect the upper limbs and typically overlay and impede voluntary actions.

Accurate phenotyping is crucial for clinical decision-making and appropriate treatment. However, movement disorders can pose significant diagnostic challenges in clinical practice due to the overlap of their symptoms. For instance, irregular forms of tremor and high-frequency cortical myoclonus can result in very similar clinical presentations. Hence, differentiating between tremor and myoclonus remains challenging, while a quick and accurate diagnostic process would greatly benefit patients. Frequently, surface electromyography (EMG) and accelerometry (ACC) measurements are applied in clinical practice for difficult-to-classify cases. Machine learning techniques have been applied in various studies, often focusing on tremor severity assessment or subtype identification. A review of machine learning methods in tremor analysis, data modalities, feature representations, and preprocessing techniques can be found in [3].

To the best of our knowledge, a direct analysis of ACC spectra by machine learning has not been studied for the discrimination of tremor and myoclonus.

In the work presented here, we apply prototype-based systems in combination with relevance learning [4]–[7] directly to the power spectra obtained from accelerometry measurements. Preliminary studies have shown that this approach yields superior performance compared to the use of engineered features derived from the spectra. Besides excellent performance, the employed method, Generalized Matrix Relevance Learning Vector Quantization (GMLVQ), enables the meaningful interpretation of prototypes and relevances, giving insight into the characteristic disease patterns. This explanatory power is enhanced when integrating information of different sensors and tasks per subject in the training.

The medical aspects and implications of these studies have been presented in greater detail in a recent, specialized publication [8]. In the current paper we focus on the machine learning analysis, discuss additional experiments, and extend the investigation of the feature relevances.

2 Setup and Data

Data was acquired in a cross-sectional study as part of the Next Move in Movement Disorders (NEMO) project [9, 10]. Its aim is to develop novel approaches to the phenotyping of movement disorders to support clinicians.

In this proof of concept study, we considered 19 patients with essential tremor (ET) and 19 patients with cortical myoclonus (CM). The diagnoses used as target labels in our analysis were provided by international panels of experts with very good inter-expert agreements. Accelerometry measurements were performed in one resting condition, eight postures, and twelve action tasks. In total, eight sensors were placed bilaterally on the upper arms (UA), forearms (FA), hands (HA), and index fingers (IF) of the participants. Here, we restrict the discussion to a representative selection of tasks:
(P) Posture: *pronated outstretched arms and relaxed wrists (both sides),*
(A) Action: *four-finger tapping task (right side, left side inactive),*
(R) Rest: *arms at rest, in a relaxed position on the subject's lap (both sides),*
(P*) Postures*: *combination of four postures with outstretched arms and varying hand or wrist positions (see section 4).*
For the analysis of the other tasks, see [8].

After standard preprocessing steps (see [8] and references therein) individual sensor recordings were represented by log-transformed power spectra, with the considered frequency range of 1-30 Hz being discretized into 99 bins of equal size. Additional experiments with further normalization, e.g., with respect to the L_1- or L_2-norm of the feature vectors, showed little or no effect on the structure and performance of the classifiers. In some tasks, single measurements had to be excluded due to technical problems, reducing the availability of samples from 19 to 18 in the ET cohort. For a detailed description of the patient cohorts, statistical aspects, and details of the data acquisition and processing, see [8].

Note that, for the posture (P) and resting condition (R), the spectra from all included sensors were averaged, resulting in 99-dim. feature vectors. The action task (A) involves finger tapping with the right hand only. Here, the spectra were

	no. of samples						
Task	ET	CM	UA	FA	HA	IF	UA, FA, IF
(P)	18	19	.98 (.06)	.99 (.04)	.95 (.08)	.93 (.09)	1.0 (.01)
(A)	19	19	.85 (.14)	.89 (.12)	.95 (.07)	.99 (.03)	1.0 (.02)
(R)	18	19	.57 (.17)	.71 (.18)	.61 (.20)	.54 (.23)	.76 (.20)
(P*)	75	73	.96 (.05)	.95 (.06)	.97 (.03)	.95 (.06)	.96 (.05)

Table 1: Average AUROC and standard deviations observed in 20 runs of five-fold cross-validation using individual sensors (UA: upper arm, FA: forearm, HA: hand, IF: index finger) and the combination of UA, FA, and IF, see Section 2.

averaged for each arm, and the left and right sides were concatenated, resulting in 198-dim. feature vectors. We only present results for the combination of UA, FA, and IF sensors; see [8] for the results of other subsets.

3 Machine learning analysis: GMLVQ

Prototype-based learning systems such as GMLVQ are inherently interpretable [4]–[7]. A GMLVQ training process optimizes a number of representative prototypes per class, as well as, in this case, one global relevance matrix Λ. The latter defines a generalized quadratic distance of the form

$$d(\mathbf{w}_j, \mathbf{x}) = (\mathbf{x} - \mathbf{w}_j)^\top \Lambda (\mathbf{x} - \mathbf{w}_j) \tag{1}$$

between a prototoype vector $\mathbf{w}_j \in \mathbb{R}^d$ and a data point $\mathbf{x} \in \mathbb{R}^d$. The adaptive, positive semi-definite relevance matrix is conveniently re-parameterized as $\Lambda = \Omega^\top \Omega$ with the auxiliary $\Omega \in \mathbb{R}^{d \times d}$. During training, the auxiliary matrix Ω is adapted.

In the classification problem we consider feature vectors $\mathbf{x}$ representing discretized power spectra as explained in the previous section. Hence, the outcome of the training is a set of prototypes, which can be interpreted as representative power spectra and a discriminative distance measure parameterized by $\Lambda \in \mathbb{R}^{d \times d}$. In the working phase, spectra are assigned to the class represented by the nearest of the prototypes. Here, we focused on binary classification problems and used only one prototype per class.

Given a labeled set of feature vectors $\{\mathbf{x}^\mu\}_{\mu=1}^P$, the prototypes and matrix Ω are optimized w.r.t. the cost function [6, 7]

$$E = \textstyle\sum_{\mu=1}^P \phi \left[\frac{d(\mathbf{w}_+, \mathbf{x}^\mu) - d(\mathbf{w}_-, \mathbf{x}^\mu)}{d(\mathbf{w}_+, \mathbf{x}^\mu) + d(\mathbf{w}_-, \mathbf{x}^\mu)} \right], \quad \text{with } \phi(z) = z \text{ in this work.} \tag{2}$$

Here, $\mathbf{w}_+$ denotes the closest prototype of the *same* class as $\mathbf{x}^\mu$, and $\mathbf{w}_-$ the closest prototype representing a *different* class. In addition, Ω is normalized after each update as $\sum_{i,j} \Omega_{ij}^2 = \sum_i \Lambda_{ii} = 1$ to improve numerical stability and achieve comparability of individual matrices.

We employed the python scikit-learn compatible package sklvq [11] with the following settings: adaptive squared Euclidean distance, identity function as activation, and waypoint gradient decent (wgd) as solver with averages computed

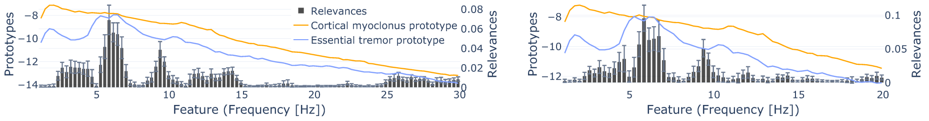

Fig. 1: Averaged results of 100 training runs, task (P), UA, FA, and IF, based on frequencies 1-30 Hz (left) and 1-20 Hz (right). The curves represent the prototypes (blue: ET, yellow: CM), while the bars correspond to the diagonal elements of Λ in the considered frequency range.

over five updates and a total number of 50 wgd steps. Initial step sizes were set to 2.0 and 1.0 for prototype and matrix updates, respectively.

All reported results were obtained in twenty randomized repetitions of five-fold cross-validation, corresponding to 100 training runs per setting. In each run, training and validation data were z-scored using means and standard deviations of the training set. Performance was evaluated in terms of the Area under the ROC (AUROC). In addition, we present relevances and prototypes averaged over all 100 runs.

4 Results and Discussion

Classification performance: In general, the GMLVQ classifiers achieved very good to excellent validation performance, see [8] for detailed considerations. Table 1 shows the results for the selected example of a posture (P) and an action task (A) for different sensors and the combination thereof with near perfect discrimination (mean AUROC=1.0, with standard deviations (SD) 0.01 in (P), and 0.02 in (A)). In the resting condition (R), the influence of CM or ET is expected to be very limited as both disorders generically exacerbate during voluntary action. Reassuringly, the attempt to distinguish the disorders in rest essentially fails and yields only relatively poor accuracies.

As confirmed in [8], performances vary slightly with the tasks and sensors considered, but the results do not suggest clear preferential setups. Where available, several body segments should be integrated into the analysis to achieve a robust, reliable performance. However, the results of our proof of concept study show that even single sensors and specific pairs of muscles might suffice to successfully discriminate ET from CM cases.

Prototypes and Relevances: Fig. 1 (left) displays example prototypes and diagonal relevances for the classification of samples in task (P), using the averaged spectra from the UA, FA, and IF sensors. We observe that a few frequency ranges play an important role. Most prominently, the diagonal relevances display a pronounced peak in the range of ca. 5-7Hz. In this range, the ET prototype also displays a peak that is absent in the CM prototype. The latter displays a much broader spread of power, while the peak of the ET prototype is accompanied by local minima around 2Hz and 9Hz. The high relevances at these frequencies

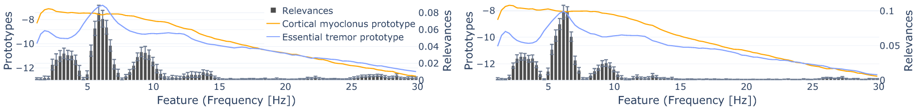

Fig. 2: Averaged results of 100 training runs, combined tasks (P*) as described in Sec. 1, averaged over UA, FA, and IF (left), and sensors presented as separate samples (right). Details as in Fig. 1.

indicate that the peak structure is an important discriminative property of the spectra, significantly determining the classification. In the literature, frequencies between 4 and 8 Hz are reported as typical tremor frequencies for ET [12].

The relevance profile shown in Fig. 1 (left) suggests a moderate yet significant importance of the high frequency range 25-30 Hz. This appeared rather implausible, an impression which was confirmed when restricting the feature vectors and prototypes to the range of, say, 1-20 Hz. Then, a similar accumulation of apparent relevance near 20 Hz was observed, see Fig. 1 (right). In order to test our hypothesis that the effect is due to the relatively small sample size and the noise associated with it, we performed modified training processes by considering spectra from qualitatively similar tasks for each patient. For instance, in addition to (P) we considered three more tasks involving *outstretched arms* but with different static hand or wrist positions. This way, the number of available spectra increases per patient and we refer to this data set as (P*). Assuming equivalent characteristics, we also determined the performance and relevance profile of the corresponding GMLVQ classifier in twenty random repetitions of five-fold patient-wise cross validation. Retaining near-perfect performance (AUROC (SD): 0.96 (0.05)), we now observed significantly reduced relevances in the high-frequency range, see Fig. 2 (left). We further increased the number of available spectra per patient by presenting sensor pairs as separate samples, yielding excellent performance (AUROC (SD): 0.97 (0.04)) and an even clearer relevance profile as shown in Fig. 2 (right). The increased number of samples per patient also enables us to consider a majority vote across all predictions, accompanied by a certainty score in terms of the percentage of correctly classified samples for each patient. Here, the majority vote achieves an accuracy (SD) of 0.95 (0.07) with 36 patients having a certainty score above 80% and only 2 patients below

Sensor	ACC (SD)
UA	.93 (.18)
FA	.95 (.15)
HA	.94 (.15)
IF	.91 (.19)

Task	ACC (SD)
Supinated arms and hands	.89 (.32)
Pronated arms relaxed wrists	.93 (.15)
Pronated arms and hands	.95 (.18)
Pronated arms extended wrists	.91 (.25)

Table 2: Average accuracy and SD per sensor (left) and task (right) observed in 20 runs of five-fold cross-validation presenting power spectra of tasks (P*) and sensors (UA, FA, HA, IF) as individual samples.

this threshold. Additionally, it provides insight into the prediction potential of tasks and sensor placements, represented by task-wise or sensor-wise accuracy, which is crucial for selecting suitable setups in clinical applications; see Table 2.

5 Summary and Outlook

In summary, our proof of concept study shows that a reliable discrimination of ET and CM should be possible on the basis of power spectra derived from accelerometric measurements. The classification accuracy appeared to be robust against the specific choice of postures and dynamic tasks. Similar performances were also observed for the use of various sensor locations or combinations thereof. The interpretation of GMLVQ feature relevances shows that the shape of the spectra relative to the typical tremor fequencies bears great importance.

The relatively small number of subjects clearly constitutes a limitation of the study. Future investigations would benefit from larger, potentially prospective, patient cohorts. Another important goal is the inclusion of additional hyperkinetic movement disorders, such as dystonia and functional disorders. It will require extensions of the model framework to different setups and modalities.

References

[1] K.P. Bhatia, P. Bain, N. Bajaj et al. Consensus Statement on the classification of tremors. From the task force on tremor of the International Parkinson and Movement Disorder Society. Movement Disorders 33(1): 75-87, 2018.

[2] R. Zutt, M.E. van Egmond, J.W. Elting et al. A novel diagnostic approach to patients with myoclonus. Nature Reviews Neurolology 11: 687-697, 2015.

[3] A. De, K.P. Bhatia, J. Volkmann et al. Machine Learning in Tremor Analysis: Critique and Directions. Mov Disord, 38: 717-731, 2023.

[4] T. Kohonen. Self-Organizing Maps. Springer, Berlin, 1997.

[5] D. Nova and P.A. Estévez. A review of Learning Vector Quantization Classifiers. Neural Computing and Applications, 25: 511-524, 2014.

[6] A.S. Sato and K. Yamada. Generalized Learning Vector Quantization. In *Advances in Neural Information Processing Systems* 7: 423–429, 1995.

[7] P. Schneider, M. Biehl and B. Hammer. Adaptive relevance matrices in Learning Vector Quantization. *Neural Computation*, 21(12):3532-3561, 2009.

[8] E.L. van den Brandhof, I. Tuitert, A.M.M. van der Stouwe et al. Explainable machine learning for movement disorders - classification of tremor and myoclonus. Computers in Biology and Medicine, in press, 2024.

[9] A.M.M. van der Stouwe et al. Next move in movement disorders (NEMO): Developing a computer-aided classification tool for hyperkinetic movement disorders. BMJ Open 11 (10), 2021.

[10] J.R. Dalenberg, D.E. Peretti, L.R. Marapin et al. Next move in movement disorders: neuroimaging protocols for hyperkinetic movement disorders. Frontiers in human neuroscience, 18, 1406786, 2024.

[11] R. van Veen, M. Biehl, and G. J. de Vries. sklvq: Scikit learning vector quantization. Journal of Machine Learning Research 22(231): 1-6, 2021.

[12] A.M.M. van der Stouwe, J.W. Elting, J.H. Van der Hoeven et al. How typical are 'typical' tremor characteristics? Sensitivity and specificity of five tremor phenomena. Parkinsonism Relat. Disord. 30: 23-28, 2016.