Machine Learning in Tremor Analysis: Critique and Directions

Anwesan De,¹,² Kailash P. Bhatia, MD, FRCP,³ Jens Volkmann, MD, PhD, FEAN,¹ Robert Peach, PhD,¹,⁴ and Sebastian R. Schreglmann, MD, PhD, FEBN* ⁴

¹Department of Neurology, University Hospital Wuerzburg, Wuerzburg, Germany
²Department of Electronics and Instrumentation, Birla Institute of Technology and Science, Pilani, Hyderabad, India
³Department of Clinical and Movement Neurosciences, Institute of Neurology, UCL, London, United Kingdom
⁴Department of Brain Sciences, Imperial College London, London, United Kingdom

ABSTRACT: Tremor is the most frequent human movement disorder, and its diagnosis is based on clinical assessment. Yet finding the accurate clinical diagnosis is not always straightforward. Fine-tuning of clinical diagnostic criteria over the past few decades, as well as device-based qualitative analysis, has resulted in incremental improvements to diagnostic accuracy. Accelerometric assessments are commonplace, enabling clinicians to capture high-resolution oscillatory properties of tremor, which recently have been the focus of various machine-learning (ML) studies. In this context, the application of ML models to accelerometric recordings provides the potential for less-biased classification and quantification of tremor disorders. However, if implemented incorrectly, ML can result in spurious or nongeneralizable results and misguided conclusions. This work summarizes and highlights recent developments in ML tools for tremor research, with a focus on supervised ML. We aim to highlight the opportunities and limitations of such approaches and provide future directions while simultaneously guiding the reader through the process of applying ML to analyze tremor data. We identify the need for the movement disorder community to take a more proactive role in the application of these novel analytical technologies, which so far have been predominantly pursued by the engineering and data analysis field. Ultimately, big-data approaches offer the possibility to identify generalizable patterns but warrant meaningful translation into clinical practice. © 2023 The Authors. Movement Disorders published by Wiley Periodicals LLC on behalf of International Parkinson and Movement Disorder Society.

Key Words: accelerometer; artificial intelligence; classification; feature based

Introduction

Tremor is the most common disturbance of movement in man, defined as an involuntary rhythmic, oscillating movement of a body part. Oscillatory movements are a function of a mechanical component, that is, the inherent mechanical propensity of an object to oscillate, and a central component, that is, the result of an activation of agonist–antagonist muscles by rhythmic central nervous system activity.¹ Clinically, tremor presents with a particular body distribution (affecting limbs, the head, neck, jaw, vocal cords, or palate), activation condition (postural, kinetic, intention, or task specific), and frequency. It can occur in isolation or along various additional symptoms as part of a clinical syndrome, as summarized in the most recent consensus criteria.² To date, tremor disorders remain as clinical diagnoses, and their definition has evolved with the first consensus criteria formalized only in 1998³ and most
recently updated in 2018.\textsuperscript{2} Over time, changes particularly affected the concept of essential tremor (ET),\textsuperscript{4-6} which led to the identification of specific tremor entities previously subsumed within ET. The importance of exact clinical phenotyping is documented, for example, by the fact that evidence-based treatment options greatly differ between ET and dystonic tremor (DT),\textsuperscript{9} with accelerometry by far the most likely as it relies on carefully calibrated and limiting its general use.

Correct diagnosis remains a challenge, with misdiagnosis rates reported up to 37%\textsuperscript{10} or even 50%.\textsuperscript{11} This uncertainty documents the need for more objective measures for tremor classification. In contrast to the perception of clinical raters, measurement devices provide objective measures of tremor movement, from which features can be extracted. In the ML context, a feature can be any quantifiable signal characteristic that can be fed into an algorithm. Several techniques equally fulfill quality criteria for the quantification and characterization of tremor,\textsuperscript{12} with accelerometry by far the most widely used method.

For differentiation, simple tremor metrics have consistently proven unreliable,\textsuperscript{13} and only two methods are validated to differentiate between ET and Parkinson’s disease (PD): the tremor stability index (TSI)\textsuperscript{14} was developed based on an exploratory data set of 16 PD rest tremor and 20 ET postural tremor recordings. In a validation data set of 55 PD and ET patients, a TSI cutoff of 1.05 (below: PD, above: ET) achieved a sensitivity of 95% each. However, absolute mean TSI values for ET postural tremor have been reported below 0.5 by other groups,\textsuperscript{15,16} casting doubt on its generalizability. The mean harmonic power\textsuperscript{17} was established based on 30-s postural accelerometric recordings in 41 ET and 39 PD patients, reaching a classification accuracy of 94%. In a validation data set of 41 tremor patients, it has been shown to differentiate the two entities with 90.1% sensitivity and 100% specificity.\textsuperscript{18} However, it has not been used beyond a third data set, where it was marginally outperformed by the TSI,\textsuperscript{14} most likely as it relies on carefully calibrated accelerometers,\textsuperscript{17,18} limiting its general use.

Currently, the most reliable diagnostic tests for ET,\textsuperscript{19} enhanced physiological tremor\textsuperscript{20} and functional tremor\textsuperscript{20-23} are diagnostic tools, based on the combination of electrophysiological features and clinical scoring systems,\textsuperscript{13} which again are not regularly used in clinical practice.

Meanwhile, machine-learning (ML) approaches, the cornerstone of the current artificial intelligence (AI) revolution, are combining the theoretical attraction of purely data-driven analyses and the statistical power of large data sets, revolutionizing medicine\textsuperscript{24} and neurology\textsuperscript{25} at a rapid pace. Based on the integration of multiple high-dimensional sources of data, ML can help to identify unifying, consistent, and generalizable disease characteristics.\textsuperscript{26} The promise of better detection, prediction, and treatment of human disease has been showcased in numerous examples.\textsuperscript{27-31}

The unifying principle of ML approaches is to “learn” patterns from data without human instruction. Given its analytical power and unbiased nature, ML therefore holds great potential to aid tremor research. This review summarizes the most recent developments in the application of supervised ML to tremor disorders, as well as challenges in applying and translating these exciting possibilities into clinical practice.

Literature Search Strategy and Results

PubMed searches were conducted on November 25, 2022 (date of final search), without restrictions of publication language or type, considering publications from January 1, 2009, onward and the following search terms ("tremor*" OR "tremor" OR "tremor disorder") AND ("accelerometry" OR "accelerometer") AND ("machine*" OR "machine learning"). Additional publications were identified from the reference lists of selected papers. Identified abstracts were screened and selected based on reporting human clinical data.

After removing duplicates, the described searches provided 36 articles, including data of n = 1558 participants. The majority of publications (29, n = 1059 patients) related to the detection and quantification of PD tremor, five (n = 187 patients) focused on ET, four on the differentiation between PD and ET, and one each on the quantification of physiological and fatigue-induced enhanced physiological tremor, respectively. The sample size ranged between n = 398\textsuperscript{32} and n = 1.\textsuperscript{33} The majority of studies were conducted under laboratory conditions based on “scripted protocols” with fixed recording length and positions, whereas nine studies were “unscripted,” that is, recording patients during activities of daily living (ADL; see Table 1).

Data Collection

Clinical Aspects

It is known from the classical tremor analysis literature that tremor depends not only on cause but also on limb position, activation, vigilance, and treatment state, which by nature of the experiment are controlled only under scripted protocols. In addition to the exact definition of diagnostic criteria applied, documentation of recording conditions therefore is crucial and should follow standard practice.

Recordings should therefore be performed with both forearms fully supported on an armrest to isolate the limb movement from external factors.\textsuperscript{68} As governed by clinical phenomenology,\textsuperscript{2} tremor movements are recorded in several defined positions. Although the main tremor positions are generally accepted, their
<table>
<thead>
<tr>
<th>Author</th>
<th>Sample size</th>
<th>Diagnostic criteria, clinical details given (y/n)</th>
<th>Sensors</th>
<th>Sensor position</th>
<th>Accelerometer type</th>
<th>Sampling frequency</th>
<th>Analysis method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scripted protocols</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Balachandar et al</td>
<td>78 (19 PD, 42 ET, 17 DT)</td>
<td>y, y</td>
<td>Smartphone</td>
<td>Wrist</td>
<td>Triaxial</td>
<td>50 Hz</td>
<td>RF, UMAP</td>
</tr>
<tr>
<td>Xing et al</td>
<td>398 (257 PD, 141 ET)</td>
<td>n, n</td>
<td>#317A Noraxon</td>
<td>Distal phalanx</td>
<td>Triaxial</td>
<td>12 kHz</td>
<td>NN, CNN, logistic regression, XGBoost, RF, SVM</td>
</tr>
<tr>
<td>Ali et al</td>
<td>35 (17 ET, 18 Ctrl)</td>
<td>y, y</td>
<td>IMU sensors</td>
<td>Dorsum of hand, posterior forearm, posterior upper arm</td>
<td>Triaxial</td>
<td>148.1 Hz</td>
<td>SVM, regression</td>
</tr>
<tr>
<td>Channa et al</td>
<td>40 (20 PD, 20 Ctrl)</td>
<td>n, y</td>
<td>AWEAR bracelet</td>
<td>Wrist</td>
<td>Triaxial</td>
<td>100 Hz</td>
<td>Neural net clustering, kNN</td>
</tr>
<tr>
<td>Sigcha et al</td>
<td>18 PD</td>
<td>y, y</td>
<td>Custom-built smartwatch</td>
<td>Wrist</td>
<td>Triaxial</td>
<td>50 Hz</td>
<td>Multitask classification model (CNN)</td>
</tr>
<tr>
<td>Shahtalebi et al</td>
<td>81 (47 PD, 34 ET)</td>
<td>n, y</td>
<td>#317A Noraxon</td>
<td>Dorsal middle metatarsal</td>
<td>Triaxial</td>
<td>1.5 kHz</td>
<td>Pathological hand tremor net: gated recurrent unit–based algorithm</td>
</tr>
<tr>
<td>Shahtalebi et al</td>
<td>81 (47 PD, 34 ET)</td>
<td>n, y</td>
<td>#317A Noraxon</td>
<td>Dorsal middle metatarsal</td>
<td>Triaxial</td>
<td>1.5 kHz</td>
<td>NeurDNet (ensemble architecture introduced in the paper)</td>
</tr>
<tr>
<td>Sajal et al</td>
<td>52 PD</td>
<td>n, n</td>
<td>Smartphone</td>
<td>Dorsal metatarsals</td>
<td>na</td>
<td>na</td>
<td>kNN, SVM, Naive Bayes</td>
</tr>
<tr>
<td>Zhang et al</td>
<td>12 PD</td>
<td>n, n</td>
<td>Activity AX3</td>
<td>Wrist</td>
<td>Triaxial</td>
<td>100 Hz</td>
<td>Linear SVM</td>
</tr>
<tr>
<td>de Arzájo et al (2020)</td>
<td>50 (18 PD, 32 Ctrl)</td>
<td>y, y</td>
<td>MetaMotionC</td>
<td>Dorsal middle metatarsal</td>
<td>Triaxial</td>
<td>100 Hz</td>
<td>kNN, logistic regression, SVM, linear discriminant, RF, decision tree, Gaussian naive Bayes</td>
</tr>
</tbody>
</table>

(Continues)
TABLE 1  
Continued

<table>
<thead>
<tr>
<th>Author</th>
<th>Sample size</th>
<th>Diagnostic criteria, clinical details given (y/n)</th>
<th>Sensors</th>
<th>Sensor position</th>
<th>Accelerometer type</th>
<th>Sampling frequency</th>
<th>Analysis method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aljihmani et al&lt;sup&gt;42&lt;/sup&gt;</td>
<td>40 (20 type 1 DM, 20 Ctrl)</td>
<td>n, n</td>
<td>ADXL355</td>
<td>Dorsal middle phalanx</td>
<td>Triaxial</td>
<td>45 Hz</td>
<td>Decision tree, kNN, SVM, ensemble classifier</td>
</tr>
<tr>
<td>Shawen et al&lt;sup&gt;43&lt;/sup&gt;</td>
<td>13 PD</td>
<td>n, y</td>
<td>BiostampRC + Apple watch</td>
<td>Wrist + dorsal middle metatarsal</td>
<td>Triaxial</td>
<td>62.5 + 50 Hz</td>
<td>RF</td>
</tr>
<tr>
<td>Mahadevan et al&lt;sup&gt;44&lt;/sup&gt;</td>
<td>81 (31 PD, 50 Ctrl)</td>
<td>n, y</td>
<td>Opal and BioStamp watch</td>
<td>Wrist</td>
<td>Triaxial</td>
<td>128 Hz</td>
<td>RF</td>
</tr>
<tr>
<td>Ribeiro et al&lt;sup&gt;45&lt;/sup&gt;</td>
<td>35 (14 PD, 21 Ctrl)</td>
<td>n, n</td>
<td>na</td>
<td>Hand</td>
<td>Triaxial</td>
<td>na</td>
<td>Bidirectional gated recurrent units—RNN</td>
</tr>
<tr>
<td>McNames et al&lt;sup&gt;46&lt;/sup&gt;</td>
<td>17 (10 PD, 7 Ctrl)</td>
<td>n, n</td>
<td>Opal</td>
<td>Wrist</td>
<td>Triaxial</td>
<td>128 Hz</td>
<td>Proprietary two-stage algorithm</td>
</tr>
<tr>
<td>Hsayera et al&lt;sup&gt;47&lt;/sup&gt;</td>
<td>24 PD</td>
<td>n, n</td>
<td>Great Lakes Neurotechnologies</td>
<td>Wrist and ankle</td>
<td>Triaxial</td>
<td>64 Hz</td>
<td>Gradient tree boosting and LSTM</td>
</tr>
<tr>
<td>Pulliam et al (2017)&lt;sup&gt;48&lt;/sup&gt;</td>
<td>13 PD</td>
<td>y, y</td>
<td>Accelerometer + gyroscope</td>
<td>Both wrists and ankles</td>
<td>Triaxial</td>
<td>na</td>
<td>Regression models</td>
</tr>
<tr>
<td>Sanchez-Perez et al&lt;sup&gt;49&lt;/sup&gt;</td>
<td>57 PD</td>
<td>n, y</td>
<td>na</td>
<td>Middle of lower arm</td>
<td>Triaxial</td>
<td>50 Hz</td>
<td>Fuzzy inference system</td>
</tr>
<tr>
<td>Kim et al&lt;sup&gt;50&lt;/sup&gt;</td>
<td>92 PD</td>
<td>n, y</td>
<td>Custom-developed devise</td>
<td>Finger + wrist</td>
<td>Triaxial</td>
<td>20 Hz</td>
<td>CNN</td>
</tr>
<tr>
<td>Jeon et al&lt;sup&gt;51&lt;/sup&gt;</td>
<td>85 PD</td>
<td>y, n</td>
<td>Wristwatch</td>
<td>Finger + wrist</td>
<td>Triaxial</td>
<td>125 Hz</td>
<td>PCA, kNN, SVM, discriminant analysis, decision tree</td>
</tr>
<tr>
<td>Ghassemi et al&lt;sup&gt;52&lt;/sup&gt;</td>
<td>24 (13 PD, 11 ET)</td>
<td>y, y</td>
<td>na</td>
<td>Dorsal metatarsal</td>
<td>na</td>
<td>na</td>
<td>SVM</td>
</tr>
<tr>
<td>Kostikis et al&lt;sup&gt;54&lt;/sup&gt;</td>
<td>45 (25 PD, 20 Ctrl)</td>
<td>n, y</td>
<td>Smartphone iPhone</td>
<td>Dorsal middle metatarsal</td>
<td>Triaxial</td>
<td>20 Hz</td>
<td>Tree Bagger</td>
</tr>
<tr>
<td>LeMoyne et al&lt;sup&gt;33&lt;/sup&gt;</td>
<td>1 ET (deep brain stimulation On vs. Off)</td>
<td>n, n</td>
<td>Smartphone iPhone</td>
<td>Dorsal middle metatarsal</td>
<td>na</td>
<td>100 Hz</td>
<td>SVM</td>
</tr>
</tbody>
</table>

(Continues)
<table>
<thead>
<tr>
<th>Author</th>
<th>Sample size</th>
<th>Diagnostic criteria, clinical details given (y/n)</th>
<th>Sensors</th>
<th>Sensor position</th>
<th>Accelerometer type</th>
<th>Sampling frequency</th>
<th>Analysis method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ahlrichs and Samà</td>
<td>76 PD</td>
<td>y, n</td>
<td>Accelerometer + gyroscopes</td>
<td>Wrist</td>
<td>Triaxial</td>
<td>40 Hz</td>
<td>SVM with different kernels</td>
</tr>
<tr>
<td>Veluvolu et al (2011)</td>
<td>10 (5 surgeons, 5 Ctrl)</td>
<td>n, n</td>
<td>ADXL 203</td>
<td>na</td>
<td>Mono-axial</td>
<td>500 Hz</td>
<td>Multistep prediction with BMFLC and MS-AR model</td>
</tr>
<tr>
<td>Rigas et al (2012)</td>
<td>23 (18 PD, 5 Ctrl)</td>
<td>y, y</td>
<td>na</td>
<td>Wrist, ankle, sternum, waist</td>
<td>Triaxial</td>
<td>62.5 Hz</td>
<td>HMM</td>
</tr>
<tr>
<td>Patel et al</td>
<td>12 PD</td>
<td>n, y</td>
<td>Shimmer sensor</td>
<td>Middle of lower arm</td>
<td>Mono-axial</td>
<td>100 Hz</td>
<td>SVM</td>
</tr>
<tr>
<td>Unscripted protocols</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>San-Segundo et al</td>
<td>12 PD</td>
<td>n, y</td>
<td>Smartwatch</td>
<td>Wrist</td>
<td>Triaxial</td>
<td>800 Hz</td>
<td>CNN</td>
</tr>
<tr>
<td>Battista et al (2020)</td>
<td>20 PD</td>
<td>y, y</td>
<td>Custom-made</td>
<td>Wrist</td>
<td>Triaxial</td>
<td>na</td>
<td>Threshold method</td>
</tr>
<tr>
<td>Papadopoulos et al</td>
<td>37 PD</td>
<td>n, n</td>
<td>Inertial measurement units and smartphone</td>
<td>Hand</td>
<td>Triaxial</td>
<td>100 Hz</td>
<td>MIL</td>
</tr>
<tr>
<td>Zheng et al</td>
<td>8 ET (26 h per patient)</td>
<td>y, y</td>
<td>Pebble smartwatch</td>
<td>Wrist</td>
<td>Triaxial</td>
<td>25 Hz</td>
<td>FFT and clustering</td>
</tr>
<tr>
<td>Garcia-Magarino et al (2016)</td>
<td>21 (11 PD, 10 Ctrl)</td>
<td>n, y</td>
<td>Smartphone</td>
<td>Hand</td>
<td>Triaxial</td>
<td>100 Hz</td>
<td>“Shake detection”</td>
</tr>
<tr>
<td>Cole et al</td>
<td>23 (19 PD, 4 Ctrl)</td>
<td>n, y</td>
<td>na</td>
<td>Arms, legs, and sternum</td>
<td>Triaxial</td>
<td>1 kHz</td>
<td>Dynamic neural network, DSVM, HMM</td>
</tr>
<tr>
<td>Tzallas et al (2014)</td>
<td>48 PD</td>
<td>y, n</td>
<td>Custom-made</td>
<td>Wrist, ankles, and waist</td>
<td>Triaxial</td>
<td>na</td>
<td>HMM, SVM, RF</td>
</tr>
<tr>
<td>Cole et al (2012)</td>
<td>na (29 hours PD, 15h Ctrl)</td>
<td>n, n</td>
<td>na</td>
<td>One sensor per limb</td>
<td>Triaxial</td>
<td>1 kHz</td>
<td>SVM, dynamic neural network</td>
</tr>
<tr>
<td>Roy et al</td>
<td>23 (19 PD, 4 Ctrl)</td>
<td>n, y</td>
<td>Trigno Delsys</td>
<td>One sensor per limb</td>
<td>Triaxial</td>
<td>na</td>
<td>Multiple dynamic neural network classifier</td>
</tr>
</tbody>
</table>

**Abbreviations:** PD, Parkinson’s disease; ET, essential tremor; DT, dystonic tremor; RF, Random Forest; UMAP-XGBoost, Extreme Gradient Boosting; NN, neural networks; CNN, convolutional neural network; SVM, support vector machine; RNN, recurrent neural networks; DM, diabetes mellitus; kNN, k nearest neighbors; na, not available; LSTM, long short-term memory; PCA, principal component analysis; BMFLC, band-limited multiple Fourier linear combiner; MS-AR, multistep-autoregressive; HMM, hidden Markov models; MIL, multiple-instance learning; FFT, fast Fourier transform; DSVM, dynamic support vector machine.
exact execution differs from center to center and throughout the tremor literature: this pertains predominantly to rest (hands hanging freely without active holding) or resting or on the ulnar side of the hand on middle phalangeal, with heavier devices distal 34,36,37,40,44 arithmetic stress and is not commonly to com-
Another sensor characteristic not commonly to com-
In the majority of studies the sampling fre-
These aspects greatly limit the generaliz-
the ideal sensor should also or not speci-
or wrist
diff-
representing common clinical
to 60 seconds,
or own lap,
or throughout a medica-
or not

For posture recordings many authors adopted the clas-
classic clinical position (extending both arms and hands in parallel in front of the chest 32,33), as previously done in several seminal analytical tremor publications, 14,22,70–72 over the more lab-oriented routine of extending the hands while the arms are still resting on an arm-rest. 68,73,74 Others again used certain activation para-
digms, such as spreading fingers, 75,76 arithmetic stress 77 to induce tremor, or even measured patients lying supine to exclude all possible interfering body movements. 77,78

It is very likely that although clinically the activation pattern between, for example, slightly different posture positions might be fundamentally the same, the physical signal characteristics might be very different. Although the effects of slightly different accelerometer positions have been shown to influence classic signal characteristics such as amplitude, frequency, and total power, 73,79 different rest and posture recording positions have not been systematically compared so far. One argument in favor of the clinical measurement position is the more direct correlation with routine clinical diagnosis and scoring. Future studies should compare such positions, identifying the ideal tremor recording position.

It is generally accepted that recordings should be done after sufficient (ideally >12 hours for levodopa, >30 hours for dopamine agonists, >24 hours for β-blockers, and >36 hours for primidone) withdrawal of tremor-influencing medication and other substances such as nicotine and caffeine or tremor-inducing medication. 80,81

Our search identified several disadvantages in the tremor ML literature so far, as only 12 studies (of 36, representing 455 participants) stated the diagnostic criteria applied, and only 22 stated the summary clinical/demographic details. The patient treatment state was reported in some papers 33,43,47,53–55,58,64,67; whereas some patients were recorded in their medication on state, 37,62 other reports included patients on various medication states, 32,49 or throughout a medication cycle. 64,67 These aspects greatly limit the generalizability of results, as the clinical context is paramount for interpretation.

Experimental Setup

The exact placement of sensors does influence acceler-
ometro signal characteristics. 73,79 Sensor placement in the aforementioned studies was not uniform, ranging from, for example, wrist, 34,36,37,40,44 dorsal middle metacarpal, 38,39,52,54,82 middle phalangeal, 42 distal phalangeal, 32 and middle of the lower arm 49,58 to combinations of, for example, wrist and finger 43,51 or wrist and ankle. 47 Traditionally, sensors have been—with some center-to-center differences—placed on the back of the hand. 68,73,74 Studies assessing the ideal sensor placement for tremor analysis showed contrasting results. 35,83

Recording Device

Today capacitive microelectromechanical accelerometers—measuring translational acceleration—are most widely used. 84,85 To adequately capture the entire dynamics of the tremor signal, the device sampling rate, frequency range, and sensitivity have to be optimized for the respective signal. Most modern digital transducers have their own built-in A/D (analog-to-digital) converter or use pulse-width modulation to estimate the width of the pulse generated proportional to the physical quantity being measured. As a rule, the A/D converter must have a sampling rate of at least twice—ideally four times—the highest frequency of interest (Nyquist frequency) to avoid aliasing artifacts. 12,68,85 In the majority of studies the sampling frequency was 100 Hz, for example, 33,39,40,51,58 whereas some authors recorded at frequencies far above four times, 32,38,82 respectively, just about two-times the Nyquist frequency. 54

Gravitational artifacts could be overcome by the application of multiple accelerometers on the same limb. 56 In practice, this has been applied very rarely 87 and is not commonly recommended for clinical or research recordings. 68

It is encouraging that most of the aforementioned studies used triaxial recordings, which, by the nature of providing a vector sum, are independent of the main movement axis. It, however, remains to be proven if recording tremor along three axes is superior to mono-axial recordings. 68

Adding gyroscope data to triaxial recordings, however, has been shown not to increase tremor scoring accuracy. 43

The accelerometer sensor range, that is, the maximal acceleration that can be recorded, ranged from 2 g 37 to 16 g. 51 Another sensor characteristic not commonly reported in the literature is the sensor sensitivity, measured in mV/g, governing the resolution at which a sensor can measure acceleration. This usually ranges between 20 and 100 mV/g, with exceptional devices providing a resolution of up to 800 mV/G. As the natural frequency of a limb depends on its weight, 1 the ideal sensor should also be as small and light as possible, 85 with heavier devices (eg, smartphones) presumably interfering more.

Signal Length and Preprocessing

As tremor intensity physiologically fluctuates with time, the recording length governs which temporal aspects of the signal are included in the analysis. For most studies, the length per tremor signal recorded ranged from 20 to 30 32 to 60 seconds, 51 representing common clinical examination durations. Whereas some studies suggested that differentiation accuracy plateaus with recording lengths from 5 seconds upward, 38 other studies for tremor
detection “in-the-wild” analyzed data collected over up to 26 hours per patient. To avoid biasing the analysis and learning features dependent on time-series length, equal lengths/equal number of data points of signal from all groups/conditions should be entered into the analytical pipeline. For in-the-wild experiments, it is necessary to segment the unequal prolonged time periods of experimentation into equal-length segments for effective classification. One study used a bag-of-features classification, where an overlapping window of fixed size was used to break the data into segments, and windows with the highest energy in the tremor frequency range were selected.

**ML Analysis**

ML methods can generally be separated into supervised learning, meaning that an algorithm is trained to detect patterns in data according to ground truth labels or “gold standard” samples fed during training (Fig. 1A), and unsupervised learning, meaning that an algorithm detects similarities and differences between samples without ground truth labels or training on standard samples (Fig. 1B). The identified studies ultimately implemented supervised learning (although unsupervised ML, such as dimensionality reduction, was sometimes used as a precursor step). We therefore have not specifically included unsupervised learning in this review, which can provide additional insights into tremor data.

**Data Preparation**

Data preparation involves the transformation of raw data into an appropriate format for modeling. First, recordings (Fig. 2A) need to be “cleaned,” removing mistakes, artifacts, or recording errors in the raw data when visually inspected in the time and frequency domain (Fig. 2B). In the absence of time stamps, as a reference for subselecting parts of the full recordings, the time window can be treated as a hyperparameter that needs to be optimized during the validation step of the ML process. However, we emphasize that biased subselecting of data, akin to “cherry picking,” may omit the physiological fluctuations inherent to tremor. Some authors clip the first and last 10 seconds of the recordings to exclude artifacts and instability associated with starting the experiment.
FIG. 2. Machine-learning workflow for analyzing tremor accelerometry recordings. (A) Raw tremor accelerometry recordings are (B) first preprocessed to clip periods of time with artifacts in both time (i) and frequency domain (ii) and then band-pass filtered to remove high- and low-frequency components. (C) Time-series features are extracted from the recordings, including (i) traditional hypothesis-driven features that are often derived from Fourier or Hilbert transforms and (ii) nondomain specific data-driven features, and which are then entered into a feature matrix (iii). (D) The ML model is trained for (i) classification or (ii) regression against the dependent variable. (E) The trained model is evaluated on a test set where the predicted classes of each sample are compared against the ground-truth classes using, for example, a confusion matrix. (F) The importance of each feature in the ML model is measured using different approaches, such as the (i) magnitude of the coefficient in a linear model or (ii) the Shapley additive explanation (SHAP) value in a nonlinear model so that (G) a clinician can make informed diagnoses and decisions. [Color figure can be viewed at wileyonlinelibrary.com]
(Fig. 2Bi). Nonphysiological artifacts due to ADL, volitional movement, and spurious artifacts are further removed by band-pass filtering (Fig. 2Bi), classically in the range of 1 to 16 Hz, 0.5 to 15 Hz, or 2 to 30 Hz.

Next, the features for best characterization of tremor time series need to be selected (Fig. 2C). Here, one can take either (1) a hypothesis-driven approach using engineered features that one expects to be predictive or (2) a data-driven approach by utilizing nondonald specific feature extraction tools or directly learning the features using a deep-learning (DL) algorithm. The former allows to input prior knowledge into the modeling process and can often provide more interpretable insights, whereas the latter is less biased and can often identify novel, unexpected, and (potentially nonlinear) relationships.

Early attempts to analyze tremors used hypothesis-driven approaches and defined features, including position-dependent tremor peak frequency (Fig. 2Ci), patient demographics, and derived metrics to further quantify the dynamic nonlinear oscillatory characteristics of tremor time series. Cross-correlation and autocorrelation features are generally accepted as good representatives of tremor characteristics. Most of the reviewed accelerometer publications engineered 1 to 10 features from both time and frequency domains, representing amplitude and regularity, spectral power, fast Fourier transform coefficients, and spectrograms.

Transitioning from a hypothesis-driven toward a data-driven approach, a selection of studies increased the number of features that they engineered (Fig. 2Cii), including 40 features, 74 features, or up to 290. Any actively selected feature might however introduce bias and limit the scope for data exploration. Recently, an entirely data-driven approach employing highly comparative time-series analysis for massive feature extraction was applied to tremor signals and was able to accurately predict between pre- and poststimulation, highlighting the potential of training predictive models of tremor signals without any domain knowledge. Finally, DL methods circumvent feature extraction steps by learning hierarchical features directly from time series. Whereas classical ML approaches allow the input of human intuition and domain knowledge, DL methods can identify complex nonlinear patterns that cannot be captured by individually engineered features.

After a feature matrix has been constructed from engineered features (Fig. 2Ciii), a second round of data cleaning can be performed on the features, for example, mean imputation to fill in missing values. Commonly, features are scaled by either normalization (bounding values between 0 and 1) or standardization (zero mean and unit variance). This is to make each feature comparatively similar in magnitude, allowing ML algorithms based on gradient descent to iterate/converge more smoothly. Some authors used standardization to scale the features to ensure comparability between ML model performance, whereas others normalized, minimized to maximum, their power-spectrum density (PSD)—based features before implementing a two-stage algorithm.

Models

The choice of supervised ML model is partially dependent on the research question and the accompanying data. Generally, supervised ML models can be separated into regression and classification models (Fig. 2D), whereby the former predict continuous values (eg, a tremor score) and the latter predict discrete values (eg, disease). A wide range of linear, ensemble, and DL methods have been applied for ML-based classification, and most of the reviewed publications applied all or a combination of these.

Linear discriminant analysis (LDA) and support vector machines (SVM) are frequently used algorithms for tremor classification. LDA inherently provides dimensionality reduction while preserving the interclass variance and ensuring maximum class separability, whereas SVMs try to estimate the best hyperplane that would serve as a boundary between classes by mapping the input to a higher dimension. Other popular ML algorithms for tremor classification include Random Forest, an ensemble algorithm used for both regression and classification tasks with an additional layer of randomness for bootstrap aggregation, and naive-Bayes classifier, based on Bayes’ theorem, that assumes features are independent and each feature is learned separately, simplifying the learning in comparison to other algorithms.

A DL model employs the use of artificial neural networks (ANN) several layers deep, which provides a very powerful nonlinear architecture to analyze the input. The characteristics of the hidden layers can be easily modified, making this architecture very flexible. The neurons in the hidden layer help recognize the features of time-series data, whereas the individual weights associated with the neurons represent the feature. DL models are completely independent of user-defined features, and time-series data with minimal preprocessing can be input directly into the model.

Convolutional neural networks are a modification of ANNs that learn relational information between spatially close data, for example, across the three dimensions of triaxial data or through time. Alternatives for continuous time-series data analysis include sequence models like recurrent neural networks (RNN), which learn features across sequential timesteps, pushing the most important features from each timestep forward for a better representation of the
sequential data. RNNs and its modified versions, long short-term memory (LSTM) and gated recurrent unit, proved very useful in sequential tremor time-series analysis.\textsuperscript{45,97}

DL methods, however, have several disadvantages. They are very data intensive, and proper training requires large data sets and expensive computations. ANNs are often described as black boxes due to the long series of complex operations that are difficult to disentangle, although explainable AI is emerging as a possible route to interpretability.\textsuperscript{98} The absence of handcrafted custom features lends this approach more objectiveness but leads to a rather opaque view of features and the inner workings of the algorithms. In this context, careful clinical phenotyping of the training set becomes an even more pressing necessity.

A common approach therefore is to use a variety of models, including linear, ensemble, and DL-based algorithms.\textsuperscript{32,99} This has similarly been tried in unscripted experiments.\textsuperscript{59} An interesting study used multiple-instance learning, whereby a bag (sequence of signal segments associated with a single class label) was provided as input and the bag (ie, not singular data points) mapped to a label, overcoming noise in the data set.\textsuperscript{61}

### Training and Evaluation

Once an appropriate ML model has been chosen, it must be optimally trained and evaluated. Training involves optimizing the model parameters and hyperparameters on a training set and a validation set, respectively, using cross-validation, whereas evaluation usually requires testing the optimized model on unseen test sets.

Cross-validation is a technique of splitting the data into a specified number of folds and permuting the training and validation sets among them. This helps to tune hyperparameters, overcome instability in sampling, and test the model performance on unseen data. K-fold cross-validation (splitting the training data into, eg, 10-folds) is one of the most common forms,\textsuperscript{32,36,38,47} and leave-one-out-cross-validation implies the number of folds equals the number of samples.\textsuperscript{31–33,61} To tune the hyperparameters of a model, a popularly adopted method is grid search, testing all possible combinations of hyperparameters within a defined range of values to find the optimal combination. Cross-validation is usually carried out on the training data within the grid search loops to evaluate the optimal set of hyperparameters.\textsuperscript{58}

Given a trained model, its ability to generalize to unseen data must be evaluated. Deciding the best metrics for evaluating model performance is a critical step in the process. A poor metric choice might lead to a distorted representation of the model capabilities and characteristics; for example, a model that simply classifies every sample into the same class in the presence of a large class imbalance will inevitably lead to many positively classified samples but has little clinical use. For tremor classification, model evaluation is frequently done using a confusion matrix (Fig. 2E), which allows the calculation of precision, recall, specificity, and F1 scores.\textsuperscript{32,61,100} The F1 score appears to be a favorable metric especially in unbalanced class cases.\textsuperscript{100} For binary classification models, some studies used the receiver operating characteristic curve.\textsuperscript{32,36,38} Regression models must be evaluated using some measure of distance between the ground truth and predicted scores; metrics such as mean absolute error or root mean square have been used in this context.\textsuperscript{35,51}

The computational cost to train ML models varies considerably. We emphasize that standard ML models (linear models, SVMs, Random Forest) with a few hundred samples can easily be trained on modern laptops in a matter of seconds to minutes. The computational cost increases with larger data sets and algorithms, with increasing numbers of parameters to optimize, that is, ANN, which can take hours to days to train. For tremor measurements, the computational cost of such models will not yet be a limiting factor.

### Interpretation

Adequate interpretation of ML results in the context of each experiment is paramount for insights into disease mechanisms and clinical translation. As the tremor field transitions from hypothesis-driven analyses toward data-driven analyses with high-dimensional data sets and complex DL models, it is necessary to identify the features that contributed the most toward classification (Fig. 2F).

Model comparison allows to train multiple models on different features, and subsequently compare their performance, identifying the feature sets providing the strongest predictive power. In this way, several ML studies identified, for example, models training on power-spectrum engineered features\textsuperscript{101} or SVM models\textsuperscript{52} to perform best. The coefficients in linear ML models, for example, linear SVM or logistic regression, are often directly interpretable (Fig. 2Fi). For example, Ali et al\textsuperscript{35} found a linear relationship between Fahn–Tolosa–Marin tremor score and PSD features using a least-squares linear regression, with which they could define a model equation relating the contributions of each feature to the clinical score.\textsuperscript{35}

Interpreting results of nonlinear ML models is more difficult: model gain analysis can be used to identify the parts of the signal most relevant for classification, for example, after XGBoost (Extreme Gradient Boosting) and LSTM analysis,\textsuperscript{37} or the use of Shapley additive explanation (SHAP) values (Fig. 2Fii). Another approach generated visual explanations of tremor...
spectrograms using gradient-weighted class activation
mapping to highlight the regions of the spectrogram
most relevant for classification.38

Discussion

The results from the studies summarized earlier docu-
ment, in principle, the great potential of ML for the
study of tremor disorders, as time-series analysis has
evolved from comparing relatively simple metrics to
more sophisticated feature-based analysis. Concur-
rently, methods for measuring tremor signals are well
established, and the strengths and weaknesses of ML
analytical approaches are acknowledged. As evidenced
by the strong presence of authors from the engineering/
signal analysis field in the aforementioned publica-
tions, ML technology is now sufficiently developed and
widely available.

This contrasts with the lack of clinically well-defined
tremor cohorts in our search. As in other fields of
medicine,102 the absence of clinical and demographic
details from most of the papers summarized earlier
limits their clinical value: similar to training a clini-
cian’s eye to detect clinical patterns, ML-based analyses can
only be as good as the data fed to the algorithm. This
pertains to the meticulous clinical description to deter-
mine potential bias. Further, most analyses are limited
to historically established, well-known features, poten-
tially introducing signal bias and limiting the scope for
data exploration. Together with relatively small sample
sizes, the single-center design by its nature is limiting
the ability to identify generalizable, disease-specific
characteristics.

What Is Ground Truth in Tremor Research?

Accurate tremor classification has proven to be a
challenge and an ongoing struggle for the movement
disorder community,5,6,11,103 and several concepts con-
tinue to be debated. As tremor remains a clinical diag-
nosis without clear biomarker profile beyond DAT
SPECT (dopamine transporter-single-photon emission
computed tomography) scans to quantify dopaminergic
neurodegeneration,104 it is important to recognize that
it is the combination of diagnostic criteria and their
clinical interpretation that governs how individual
symptoms are categorized. As there is no doubt this
already affects the differentiation of well-established
concepts such as ET and PD, this is even more likely to
be the case in, for example, comparatively recent addi-
tions to the diagnostic spectrum, such as DT and
tremor associated with dystonia.2,105–107 In the absence
of consensus biomarkers, the clinical diagnosis remains
the gold standard for comparison for now. This in turn
implies that, by its nature, absolute ground truth in
tremor disorders is veiled by uncertainty.

What to Expect from ML-Based Tremor Studies

It is therefore important to correctly set the expecta-
tions for ML tremor studies. They can only learn/identify patterns laid out by the clinical diagnoses fed in
the training data set. Therefore, it is neither realistic
nor desirable to replicate prediction accuracies 100% in
mono-centric data sets but rather to identify truly
disease- and not center-, device-, or clinician-specific
characteristics. The larger the pool of recordings from
several centers and populations, the more representa-
tive will be the results. Purely data-driven attempts,34
aiming at clustering patient recordings without the in-
fuence of clinical diagnoses, might provide additional
insights, if applied to clinically well-characterized mul-
ticenter data sets of sufficient size. First, clustering exer-
cises based on a very limited number of manually
chosen variables allowed to identify patient subgroups
based on medication response.108 The power of such
approaches becomes evident when all possible move-
ment characteristics are included via unbiased feature
extraction—ML-based analyses therefore bear the real-
istic potential to predict the effect of an intervention.30

Lessons from Other Domains

Despite the large number of research papers applying
ML across various domains of medicine,109 models so
far only rarely transition into routine clinical practice
due to various conceptual and methodological
issues:28,30 first, research papers that use ML to target a
medical problem often focus on the design and develop-
ment of an ML model, followed by evaluation on a lim-
lited data set, but leave validation, diffusion, and scaling
of the model into clinical care untouched. The validity
of an ML model should, however, be assessed in differ-
ent settings and across time periods to optimize its
validity.110 Second, translation necessarily requires con-
tinued monitoring and maintenance as factors such as
data quality and population characteristics change over
time. The continued training of ML models to ensure
robust predictions in medical care is an active field of
research.111 Moreover, changes in regulatory frame-
works or definitions of diseases require the evolution or
overhaul of existing models.112 Third, the predictions
made by ML models can sometimes be difficult to inter-
pret and explain, posing relevant medicolegal implica-
tions in potentially high-risk clinical decisions, as the
parties involved need to understand the reasons for a
health-care decision. Depending on the training data set
used, ML models can also exacerbate existing racial
and socioeconomic health inequalities.113

After a first wave of ML publications, several efforts
focused on identifying and overcoming relevant disad-
vantages:114 for example, a meta-analysis on ML
approaches in imaging research identified the frequently
unmet need to compare the performance of clinical and
ML diagnostic accuracy on the same data set, as well as external validation. Despite the translational difficulties of ML models, they still provide unprecedented diagnostic and prognostic opportunities, driving a shift toward precision and personalized health care likely to accelerate further.

Like traditional statistics, ML methods are prone to being biased if not correctly implemented. Over-fitting algorithms to the specific nuances of a data set can produce a model that cannot generalize to new data. Commonly, over-fitting is overcome by splitting the data set into a training set (trains the ML model), a validation set (optimizes ML hyperparameters), and a test set (evaluates final model performance). Other routes to prevent over-fitting include early stopping criteria (DL models usually include criteria to stop the training process when the model performance is no longer improving), expanding the data set (with a larger variety of samples), and regularization (penalizing the parameters with larger coefficients to limit variance in the ML model). Similarly, underlying issues with the data set can bias the trained ML model. For example, outliers, which are samples that are unlike the rest of the data set, can heavily bias ML models. The definition of an outlier is not fixed and depends on the data set; nonetheless, outliers can be detected using visualization or statistical methods, and training multiple ML algorithms can help avoid models that are more robust to outliers.

**Future Directions for Tremor Research**

ML predominantly holds great promises for improved generalizability for tremor research (Fig. 2G). From the methodological principles and advantages summarized earlier, we identify several core points to aid future analyses.

Learning from the successful application of ML in other domains, it is clear that only collaborative analyses of recordings from different centers, including as many samples as possible and sampling the whole spectrum of presentations, will provide the opportunity to detect not patient-, center-, or population- but truly disease-specific characteristics, improving diagnosis and prognosis. Attempts in this direction failed so far, possibly due to a mixture of the aforementioned limitations.

Increasing sample size alone, however, is insufficient to improve accuracies, as data set and spectrum bias, relating to the coverage of the disease/control spectrum and its distribution within the data, are further relevant factors influencing ML performance. First, this relates to well-documented clinical inclusion criteria and excluding treatment effects. Second, this should be accounted for by ideally selecting a range of participating centers, covering general neurology outpatient clinics as well as centers with dedicated specialist movement disorder expertise.

It still remains to be seen which tremor characteristics/features are ideal to be compared across centers, as different sensor positions, recording devices, and protocols influence established metrics. Thus, scripted protocols will remain superior in addressing the pressing questions in tremor research, such as improving diagnosis, prognosis, monitoring treatment effect; predicting treatment response; and exploring disease mechanism simply by reducing the amount of noise on top of layers of physiological inter- and intraindividual variability. The combination of multicenter data sets and the use of extensive, unbiased, and automatically extracted features appears a realistic strategy to overcome the problem of multiple known and unknown confounders introduced by the aforementioned factors.

Simultaneously, the interpretability of ML results will remain key for the translation of such attempts into clinical practice. It is essential that derived results be compared against known metrics and clinically evaluated so that they remain interpretable and intuitive to clinicians.

To make a translational impact on clinical care, features identified through ML analyses should be made available to the community, so they can be applied in routine clinical accelerometer assessments.

**Acknowledgments:** S.R.S. received support from the Advanced Clinician Scientist program of the Interdisciplinary Centre for Clinical Research, Wuerzburg, Germany. He is a fellow of the Thiemann-Foundation. R. P. acknowledges funding from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) Project-ID 424778381-TRR 295. Open Access funding enabled and organized by Projekt DEAL.

**Data Availability Statement**

Data sharing not applicable - no new data generated

**References**


60. Battista L, Romaniello A. A wearable tool for selective and continuous monitoring of tremor and dyskinesia in Parkinsonian patients. Parkinsonism Relat Disord 2020;77:43


111. Lenert MC, Matheny ME, Walsh CG. Prognostic models will be victims of their own success, unless…. J Am Med Inform Assoc 2019;26(12):1645–1650.


Author Roles

A.D. executed the search and wrote the first draft of the manuscript. K.P.B. and J.V. critically reviewed the manuscript. R.P. co-supervised the project. S.R.S. conceived and supervised the project and wrote the manuscript.

Full financial disclosures of all authors (for the preceding 12 months)

A.D. and R.P. declare no conflict of interest. K.P.B. received speaker honoraria from Ipsen, Merz, and MDS; received personal compensation for scientific advisory board from Mitsubishi (Neuroderm), Jazz Pharma, and Ipsen; receives royalties from the publication of Oxford Specialist Handbook of Parkinson’s Disease and Other Movement Disorders (Oxford University Press, 2008), and Cambridge Press; and received editorial work stipend from MDS for MDCP journal; S.R.S. declares personal compensation for scientific advisory board from Elemind Inc.