# Supervised Spike Sorting Feasibility of Noisy Single-Electrode Extracellular Recordings: Systematic Study of Human C-Nociceptors recorded via Microneurography

**Alina Troglio[1,2*], Peter Konradi[3], Andrea Fiebig[1,2], Ariadna Pérez Garriga[3], Rainer Röhrig[3], James Dunham[4], Ekaterina Kutafina[5,6†], Barbara Namer[1,7†]**

[1]Research Group Neuroscience, Interdisciplinary Centre for Clinical Research (IZKF), Faculty of Medicine, RWTH Aachen University, Aachen, Germany

[2]Institute of Neurophysiology, RWTH Aachen University Hospital, Aachen, Germany

[3]Institute of Medical Informatics, Faculty of Medicine, RWTH Aachen University, Aachen, Germany

[4]Anaesthesia, Pain, and Critical Care Sciences, School of Physiology, Pharmacology & Neuroscience, University of Bristol, UK

[5]Institute for Biomedical Informatics, Faculty of Medicine, University Hospital Cologne, University of Cologne, Cologne, Germany

[6]Scientific Center for Neuropathic Pain Research Aachen, SCN[AACHEN], Uniklinik RWTH Aachen University, Aachen, Germany

[7]Department of Anesthesiology, Intensive Care, Emergency and Pain Medicine, University Hospital Würzburg, Center for Interdisciplinary Pain Medicine, Würzburg, Germany

***Correspondence:**
Alina Troglio
alina.troglio@rwth-aachen.de

[†]These authors share the last authorship.

**Keywords:** spike shape, microneurography, spike sorting, classification, pain, neuropathy

## Abstract

Sorting spikes from noisy single-channel in-vivo extracellular recordings is challenging, particularly due to the lack of ground truth data. Microneurography, an electrophysiological technique for studying peripheral sensory systems, employs experimental protocols that time-lock a subset of spikes. Stable propagation speed of nerve signals enables reliable sorting of these spikes. Leveraging this property, we established ground truth labels for data collected in two European laboratories and designed an open-source pipeline to process data across diverse hardware and software systems. Using the labels derived from the time-locked spikes, we employed a supervised approach instead of the unsupervised methods typically used in spike sorting.

We evaluated multiple low-dimensional representations of spikes and found that raw signal features consistently outperformed more complex approaches, which are effective in brain recordings. However, the choice of the optimal features remained dataset-specific, influenced by the similarity of average spike shapes and the number of fibers contributing to the signal.

Based on our findings, we recommend tailoring lightweight algorithms to individual recordings and assessing the "sortability feasibility" based on achieved accuracy and the research question before proceeding with sorting of non-time-

locked spikes. Our approach provides the foundation for further development of spike sorting algorithms in noisy extracellular recordings of neural activity.


## Author Summary

Using electrophysiological methods like microneurography, scientists can record nerve activity in humans to understand how peripheral nerves transmit sensations such as pain and itch. These recordings capture electrical signals, known as spikes, which represent nerve impulses. However, since several nerve fibers are often recorded simultaneously, the differentiation of the individual spikes, known as spike sorting, is critical for accurate analysis.

Existing methods for spike sorting in single electrode in-vivo recordings are often insufficient due to low signal-to-noise ratios and the absence of ground truth data needed for validation. In microneurography, low-frequency electrical stimulation (marking method) is used routinely to label part of the recorded spikes. We applied the marking method to create a ground truth data set for developing and validating a supervised approach for spike sorting.

Our transparent and lightweight algorithm showed promising results. Their high variability between the recordings, with a strong reversed link between the morphological similarities of the different fibers' spikes and the sorting accuracy indicated a possibility to assess the "sortability" of individual recordings by applying use-case specific thresholds. This work provides a foundation for improving spike sorting in noisy peripheral nerve recordings, helping researchers study better how the nervous system processes sensations like pain and itch.


## 1 Introduction

Electrophysiological extracellular nerve recordings allow researchers to gain insight into the peripheral and central nervous system activity. In the peripheral nervous system, these recordings can capture crucial sensory information, such as object texture perception, motor action guidance, and warnings of potential tissue-damaging conditions [1, 2]. The recorded signals often originate from multiple neurons, which are referred to as units. Accurate spike sorting is critical for analyzing the functionality of individual units. In some experimental setups, particularly for multi-channel *in-vitro* recordings, the sorting task can be efficiently handled due to the high quality of the recorded data and supplementary spatial resolution [3, 4]. However, other essential techniques, such as single-electrode *in-vivo* microneurography experiments, present significant difficulties, including low signal-to-noise ratios, activity of the subject, and lack of spatial resolution, which would provide more reliability through simultaneous recordings of spikes from multiple adjacent sites [5]. On the computational side, the lack of benchmark datasets with ground truth and variability in the experimental setups across internationally distributed labs restrict methodological development.

Various algorithms and tools have been developed to address spike sorting for extracellular recordings. For instance, Spike2 (Cambridge Electronic Design Limited) offers comprehensive data acquisition and analysis solutions, and SpikeInterface [6] provides an extensive spike sorting pipeline with several sorting algorithms. While they perform well in experiments with clean recording conditions, applying them to the important use case of noisy *in-vivo* microneurography recordings can lead to unreliable results and false conclusions about neurophysiological processes [7]. However, reliability is essential for analyzing single-neuron level discharge patterns to understand the sensory input and processes, such as synaptic transmission and signal encoding [8].

In the context of pain and itch research, these discharge patterns have been particularly underexplored. The simplistic paradigm "more spikes with higher frequency result in more pain sensation" is still the only one in active use [9]. To verify the hypothesis that different discharge patterns encode itch versus pain sensation within the same nerve fiber and to gain a deeper understanding of the underlying mechanisms of chronic pain and itch, it is essential to analyze and quantify these discharge patterns comprehensively, which requires spike sorting.

The electrophysiological technique of microneurography enables extracellular recordings from the small diameter unmyelinated nerve fibers that are crucial for signaling itch and pain. By obtaining these recordings in humans, including patients, it is possible to correlate neural responses with individual perceptions [10–12]. Microneurography captures single action potentials (spikes) from single peripheral nerve fibers. Previous studies, for example, have established a link between spontaneous activity in C-fibers and neuropathic pain in humans [13]. However, the single-contact electrode used in microneurography typically captures spikes from multiple nerve fibers within a single recording session, as C-fibers are anatomically clustered together in Remak-bundles. In addition, signal analysis is a challenge as the spikes from unmyelinated nerve fibers are small in comparison to the electrical noise.
This leads to simultaneous spike recordings from multiple nerve fibers along with noise from the subject's background physiological activity, such as spikes in sympathetic or highly temperature-sensitive nerve fibers. This combination of noise and multiple nerve fibers presents a significant challenge for spike sorting methods and, subsequently, for extracting meaningful discharge patterns.

Given the complexity of multi-fiber recordings and background noise, Forster and Handwerker have already attempted to solve these problems by introducing a spike sorting approach specifically for microneurography [14]. Their method relies on thresholding to detect spikes, which works well when the signal-to-noise ratio is high. The algorithm generates “templates” of spike waveforms, which serve as references, in two phases: initially, templates are created based on detected spikes, and subsequently, it compares all spikes to these templates for sorting. While this method allows for clear visual association of spikes and removal of artifacts, it requires manual parameter adjustment and is very sensitive to signal quality. Despite the partial automatization of this approach introduced by Turnquist et al. [5], spike sorting in microneurography remains unreliable due to the complexity of spike morphologies, as spikes from different nerve fibers often have similar shapes, leading to misclassification. This issue is further complicated by variations in spike waveforms originating from a single nerve fiber caused by noise or changes in recording conditions [5].

To ensure robust microneurography studies, the “marking method” [15] was developed. It is a special stimulation protocol that time-locks a subset of spikes via electrical stimulation. It allows the experimenter to collect information about the number of fibers in the recording and to identify single neuron firing patterns linked to electrical stimulation, providing insights into peripheral neural activity.
Specifically, the marking method leverages the consistent conduction velocity of unmyelinated nerve fibers when electrically stimulated at a low frequency, for example, 0.25 Hz (see Methods). In this

manuscript, we refer to this type of stimulation as *background stimulation*. Spike detection and sorting are based on the latency of response relative to the background stimulus. We refer to a sequence of spikes resulting from a single fiber response to the background stimulus as *track.* Typically, multiple tracks are visible when latency responses are displayed sequentially. We refer to this data representation as waterfall plots (see Methods). This facilitates the identification of C-fiber subtypes and the automatic sorting of spikes from different tracks. When additional spikes are evoked by *additional stimuli* applied between two background stimuli, the conduction speed of the nerve fiber slows down, and the latencies of subsequent spikes increase. This phenomenon is known as activity-dependent slowing (ADS) [16]. The magnitude of the slowing of speed correlates roughly with the number of previously elicited spikes. The sudden increase in response latency is called a "marking" by the *additional stimuli* on the *background stimulus*. Despite ADS, the tracks remain visible, and the spikes from the tracks can be reliably classified, allowing the study of the fiber behavior under various electrical stimulation protocols. However, to study the responses to other significant stimulus types, e.g. chemical and mechanical, as well as for a quantitative assessment of spontaneous activity observed in patients with peripheral neuropathies, an approach to classify all spikes accurately remains important.

In this work, we use the marking method to create ground truth datasets with reliable labels for each fiber, providing a guideline for testing sorting algorithms. By analyzing spike morphologies and their impact on sorting accuracy, we aim to overcome the limitations of current shape-based methods and develop strategies for improving spike sorting in microneurography that may be applicable to other neural data.

To assess the generalization capabilities of our spike sorting approach, we collected recordings from two laboratories employing different hardware and software configurations for microneurography. Due to the fact that the marking method is routinely used in microneurography, we know that partial ground truth will be available for almost all experiments. Therefore, rather than employing the traditional unsupervised methods commonly used in spike sorting [17, 18], which often rely on principal component analysis (PCA) with unsupervised clustering, we introduce, to the best of our knowledge, the first exploration of a supervised approach for microneurography data. Since we have tracked spikes with ground truth data, we can use these labeled spikes to train a classifier for supervised spike sorting. The primary challenge, however, is sorting the untracked spikes for which ground truth data is unavailable. By assessing the classifier's accuracy on the tracked spikes, we can estimate its ability to reliably sort the untracked spikes despite having ground truth only for the tracked data.

In this work, we develop a computational pipeline that compares feature sets and spike morphology representations across different experimental configurations. We consider different data representations from simple (amplitude and width), through more sophisticated (SS-SPDF methods, as presented by Caro-Martín et al. [19]), to the raw signal and apply support vector machine (SVM) classification method to the per-subject classification task. Our data infrastructure, built for harmonizing and analyzing microneurography data, includes a metadata standard [20] tailored for microneurography experiments employing odML and odML-tables [21, 22], a Python library to

export data from a data acquisition system [23], and openMNGlab [24], an open-source analytical framework (in development). This infrastructure allowed us to utilize 26 recordings from microneurography laboratories in Aachen, Germany (datasets labeled with A) and Bristol, United Kingdom (datasets labeled with B).

Our pipeline will serve to efficiently and reliably analyze rare and valuable patient-derived recordings to better understand and treat chronic pain and itch. Further, it can serve as a "guideline" for testing and adapting spike sorting methods based on spike-shape feature sets in diverse neuro-electrophysiological datasets. Additionally, the marking method provides a practical example of how to generate ground truth data with minimal human effort, offering a more efficient and reliable alternative to the typically labor-intensive manual annotation process.

# 2 Results

## 2.1 Open-source spike sorting pipeline for microneurography data

We designed and evaluated a pipeline for the automated and systematic evaluation of spike sorting in microneurographic recordings [25]. This pipeline enables subsequent comparison of various feature extraction methods, as inputs for supervised machine learning models for microneurography data.

We pre-analyze microneurography recordings using a 'waterfall' representation, which facilitates visualization of spike time alignment to tracks during low-frequency stimulation (Fig 1). This alignment occurs because C-nociceptors have a consistent conduction velocity when stimulated electrically at a low frequency. This enables the identification and labeling of spike events by individual nerve fibers with a track number. In Fig 1, two tracks, Track1 (purple) and Track2 (orange), are identified from a nerve fascicle. Spikes are extracted within a 3 ms window from the signal, and the raw signal, spike and stimulation onsets, and fiber/track labels are combined into a harmonized NIX file [26]. Data harmonization is essential because different data acquisition systems are used across the two labs. We extract various morphological feature sets from the spike waveforms. These feature sets (in short: simple, SPDF, W), summarized in the table within Fig 1 and further detailed in the Methods section, allow us to investigate the spike morphologies of individual units and determine which feature set works best for sorting.

To assess the sorting accuracy, we employ a 5-fold cross-validation approach, splitting the labeled spike data into 80% training and 20% testing sets. Evaluation metrics are averaged across all folds to provide a more general performance score.

A Support Vector Machine (SVM) with a radial basis function (RBF) kernel is used as the classifier for its transparency, computational efficiency, and low number of hyperparameters. We calculate key performance metrics for sorting evaluation, including accuracy, precision, recall, and the macro-averaged F1-score (data in Supporting Information, S1-S4 files).

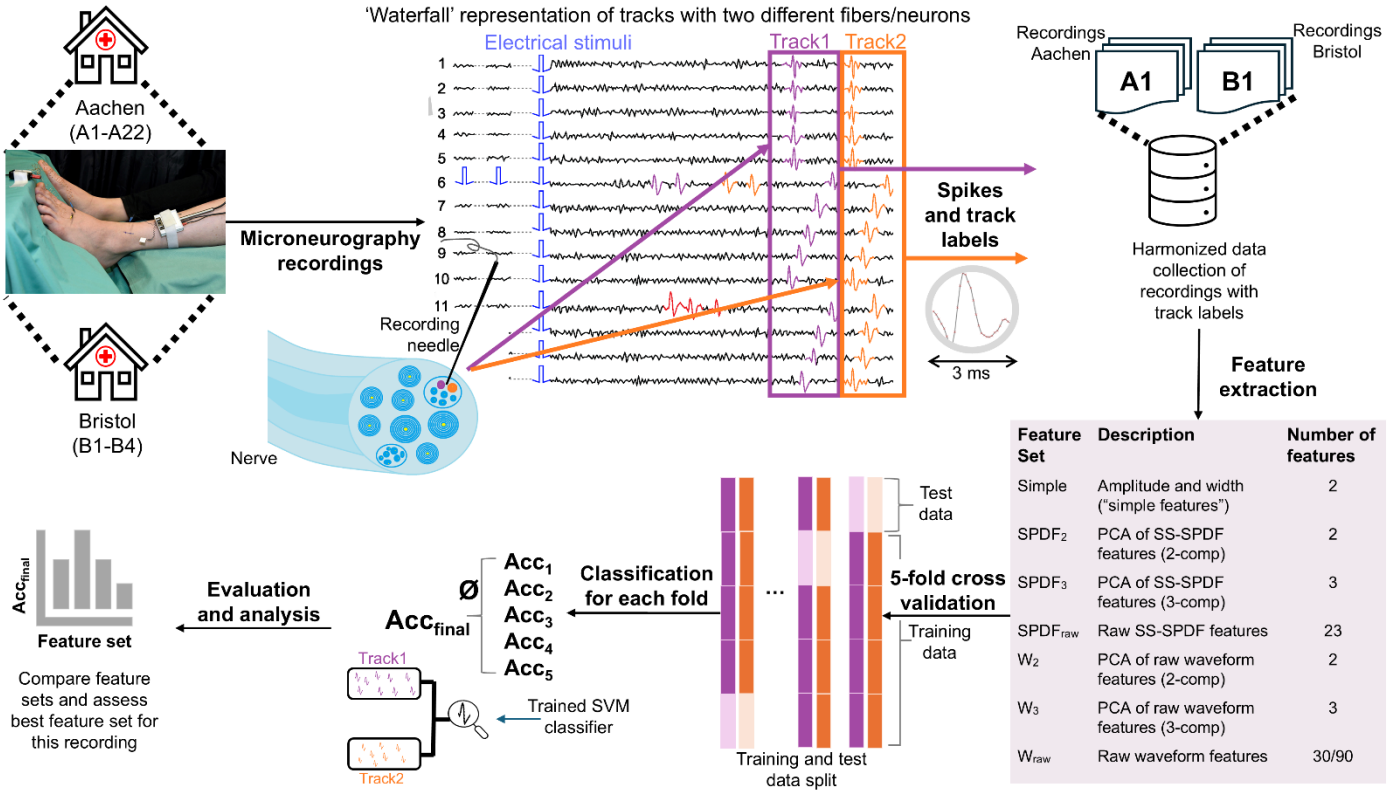


**Fig 1. Overview of the pipeline designed to sort spikes in microneurographic recordings.** Data is collected from laboratories in Aachen and Bristol and pre-analyzed using a 'waterfall' representation to align spikes based on C-fiber conduction speed. Two example *tracks* (purple and orange) are shown. The raw signal and extracted spike waveforms are harmonized, and multiple feature sets are computed (see Table) to assess their suitability for sorting. A Support Vector Machine (SVM) classifier with a radial basis function (RBF) kernel was applied to classify spikes, using 5-fold cross-validation and standard evaluation metrics (accuracy, precision, recall, and F1-score).

## 2.2. Raw spike waveform is the best input feature set for sorting spikes in microneurography recordings

We compute the averaged accuracy across all five cross-validation folds (Fig 2A). A clear trend can be observed, with the simple (amplitude and width features) and $W_2$-$W_{raw}$ (feature sets based on raw waveform data) feature sets showing superior performance compared to the other feature sets. These feature sets consistently demonstrate higher accuracy, making them more suitable for sorting. In contrast, $SPDF_2$-$SPDF_{raw}$ show median accuracies around 0.4-0.5, indicating they are ineffective. To further explore and substantiate these findings, we conducted a statistical analysis. The Wilcoxon signed-rank test with Bonferroni correction for multiple comparisons revealed significant differences between the accuracies of most feature sets (Bonferroni-adjusted threshold $\alpha_{adj} = 0.01/n \approx 0.0005$, where $n = 21$ pairwise comparisons). $W_{raw}$ performs significantly better than all other feature sets (Table 1).

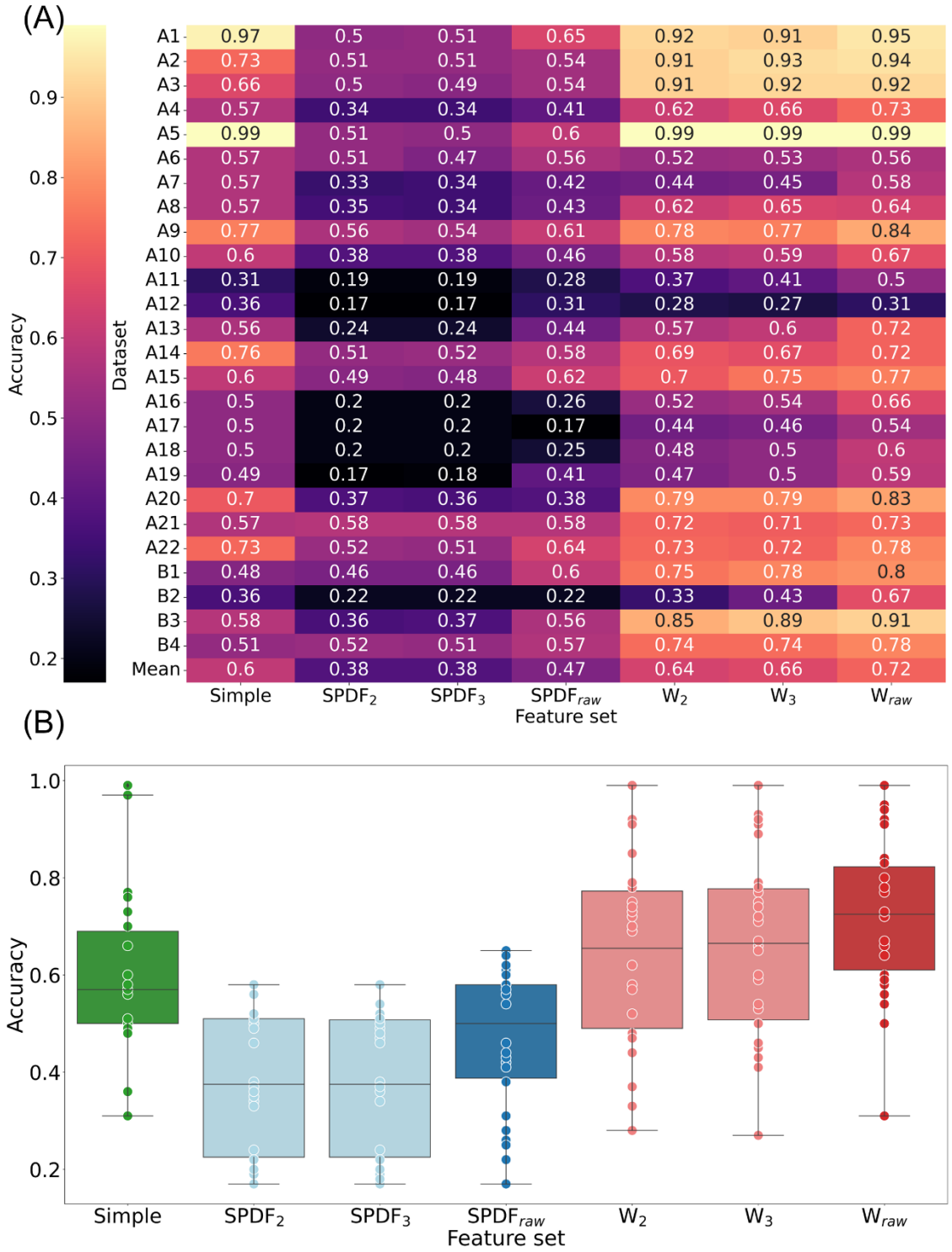


**Fig 2. Accuracy results for all datasets grouped by feature set.** (A) Individual results for each dataset. Darker shades represent lower scores, while lighter shades indicate higher scores. For $W_{raw}$, the mean accuracy is the highest, with 0.72. (B) Distribution of accuracies for each feature set. Boxes are drawn from the first quartile (median of the lower half of the score distribution) to the third quartile (median of the upper half of the score distribution). The line in the box marks the median of the scores. The lower and upper whiskers are bounded by the 1.5 interquartile range (IQR), which is the distance between the first and third quartiles. Scores outside of the 1.5 IQR bound are plotted as outliers. The dots mark the scores for individual datasets. Each color represents a different feature set group, with lighter, more transparent shades

indicating the PCA in two and three components of the corresponding feature set shown in the brighter color. Green stands for the simple feature set, blue for the feature sets of the SPDF method, and red for the raw waveform feature sets.

**Table 1.** Results of the Wilcoxon signed-rank test comparing accuracy differences across all feature sets. Each row represents a pairwise comparison between one feature set and the others, displaying the statistical significance based on the Bonferroni-corrected alpha level ($\alpha_{adj} = 0.01/21 \approx 0.0005$). Red: not significant; green: p < $\alpha_{adj}$.

| | Simple | $SPDF_2$ | $SPDF_3$ | $SPDF_{raw}$ | $W_2$ | $W_3$ | $W_{raw}$ |
|---|---|---|---|---|---|---|---|
| Simple | | | | | | | |
| $SPDF_2$ | $1.5 \times 10^{-7}$ | | | | | | |
| $SPDF_3$ | $1.5 \times 10^{-7}$ | 0.3506 | | | | | |
| $SPDF_{raw}$ | $1.597 \times 10^{-5}$ | $2.001 \times 10^{-5}$ | $2.001 \times 10^{-5}$ | | | | |
| $W_2$ | 0.1499 | $3 \times 10^{-8}$ | $3 \times 10^{-8}$ | $3 \times 10^{-7}$ | | | |
| $W_3$ | 0.0272 | $3 \times 10^{-8}$ | $3 \times 10^{-8}$ | $3 \times 10^{-7}$ | 0.0026 | | |
| $W_{raw}$ | $9.15 \times 10^{-6}$ | $3 \times 10^{-8}$ | $3 \times 10^{-8}$ | $1.5 \times 10^{-7}$ | $3 \times 10^{-8}$ | $2.1 \times 10^{-7}$ | |

Although $W_{raw}$ has the highest accuracy mean and median across all feature sets (Fig 2A and 2B), certain datasets exhibit superior performance with other feature sets, for example, dataset A14 performs best with the simple feature set as input. Therefore, evaluating the optimal feature set for each specific dataset is essential.

### 2.3. Template similarity and fiber count as indicators of sorting success

In Fig 3, we present plots for datasets A1 and A5, showing all tracked spikes and a corresponding *template* for each fiber/track in the recording, generated by averaging all spikes within each track. To quantify the impact of similar spike shapes on sorting accuracy we used three distance metrics (see Methods), including root mean square error (RMSE), to evaluate the relationship between the visual similarity of templates and its impact on sorting accuracy (Fig 4A). For instance, the RMSE between templates in dataset A1 is 1.33 (accuracy 0.97), while in dataset A6 the RMSE is significantly lower at 0.20 (accuracy 0.57) indicating higher similarity between templates and, consequently, lower sorting accuracy. This analysis demonstrates that lower distance scores often correspond to poorer sorting results due to template overlap, suggesting that template similarity is a key factor in assessing sorting quality. Additionally, distance metrics may serve as pre-sorting indicators to identify recordings that may not meet the requirements for good sorting outcomes.

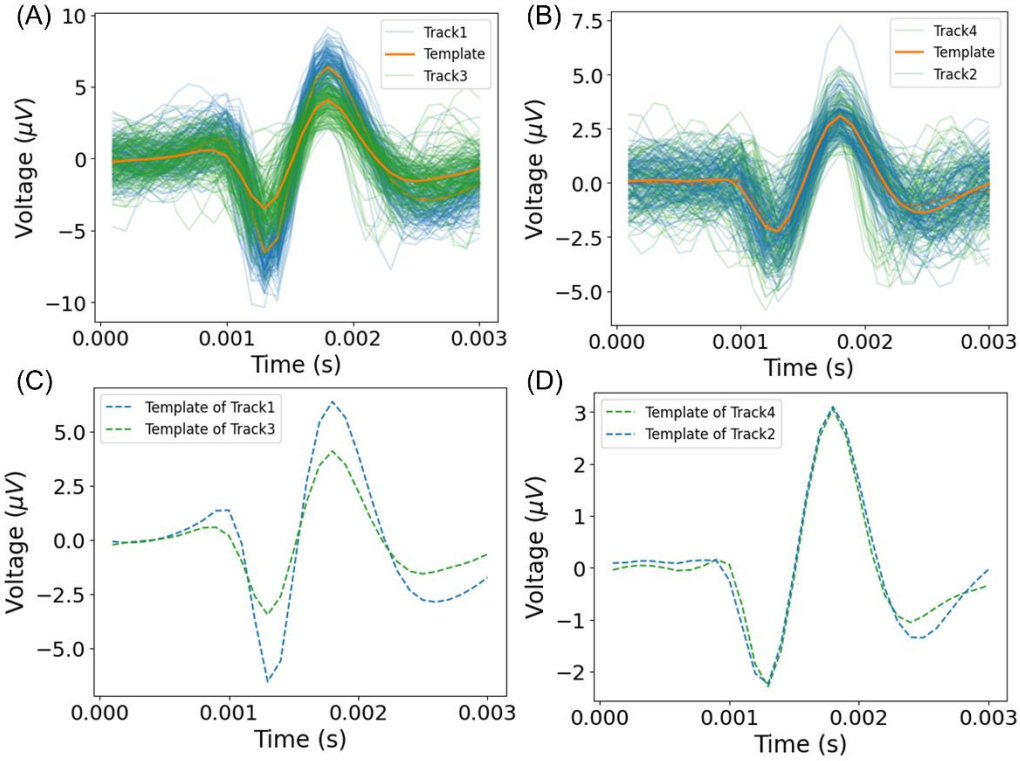


**Fig 3. Spike waveforms and corresponding templates for datasets A1 and A6, created by averaging all detected spikes.** (A) Waveforms for both tracks in recording A1, show the blue track with a higher amplitude than the green track. (B) Waveforms for both tracks in recording A6, where waveforms largely overlap except for a few outliers. (C) Templates for both tracks in recording A1 are visually distinct. (D) Templates for both tracks in recording A6, show near-complete overlap and similar amplitude, indicating high template similarity.

Fig 4B shows the mean classification accuracy across all feature sets with different marker colors representing the number of fibers or tracks (2 to 6) in each recording. As expected, recordings with fewer fibers (e.g., two fibers) consistently achieve higher classification accuracy. In contrast, the classification accuracy tends to decrease as the number of fibers increases, indicating the challenge of differentiating between more similar spikes. Beyond fiber count, the feature set choice also impacts classification performance and aligns with the previously presented results. Interestingly, feature set $SPDF_{raw}$ maintained the same mean accuracy as the raw waveform when having six fibers in a recording. This indicates that features from the SS-SPDF method might capture additional information that supports fiber classification in particularly challenging conditions.

Additionally, random chance accuracies (dashed lines) for each fiber count provide a reference for judging sorting quality. Recordings with mean accuracies close to or below these thresholds indicate poor separation of fiber classes or high template similarity. This baseline comparison allows us to identify recordings that are inherently challenging to classify accurately, even with optimal feature sets.

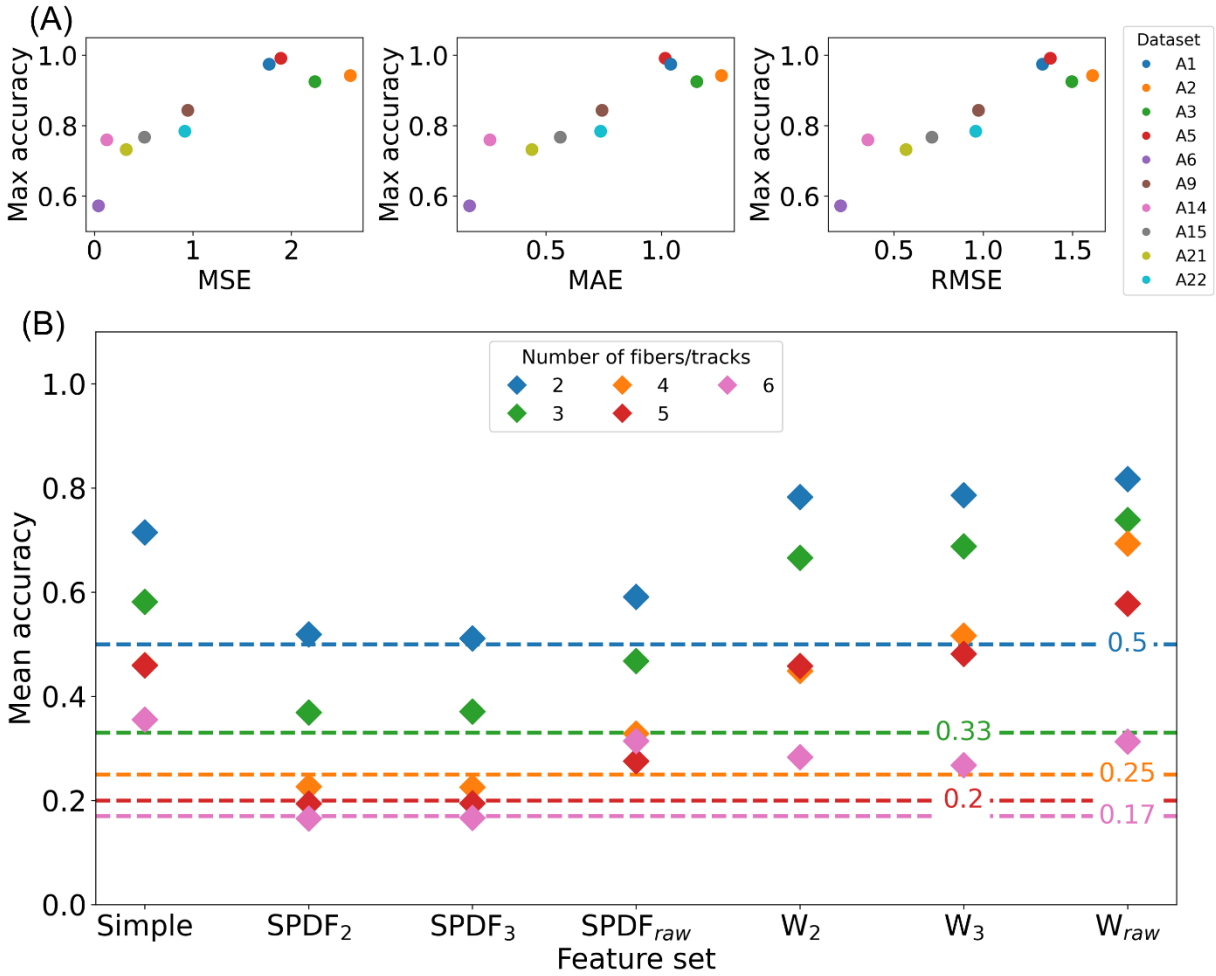


**Fig 4. Indicators of sorting success.** (A) Relationship between template distance and best-achieved sorting accuracy. The distance metrics used are the mean squared error (MSE), the mean absolute error (MAE), and the root mean squared error (RMSE). A smaller distance reflects a higher similarity between the two templates, corresponding to lower sorting accuracy. For clarity, only datasets with two fibers are considered. The color of each marker represents the value for each individual dataset. (B) Mean accuracy of each feature set (x-axis) across recordings, with accuracy values shown on the y-axis. The color of each marker represents the number of fibers tracked in the recording. The markers positioned above the horizontal line indicate that the classifier performs better than the random chance for the corresponding number of fibers.

## 2.4. Limitations of unsupervised clustering with PCA features for microneurography data

As mentioned in the introduction, the standard spike sorting process following spike detection often involves computing PCA features and examining cluster separations to distinguish fiber/track classes [18, 27]. This approach assumes that unsupervised clustering can effectively differentiate between classes by relying on the natural separation of clusters in the PCA space. However, our findings suggest this method is unsuitable for microneurography data.

In recordings with high sorting accuracy (e.g., recording A1 or A5), plotting PCA features along two principal components reveals a large, single cluster (Fig 5A) with no apparent separation between fiber classes for A1. Without prior labeling, it is not possible to distinguish distinct clusters in this recording, indicating that unsupervised clustering methods may struggle to define clear boundaries

between fiber/track classes. When labels are introduced, however, a distinct decision boundary becomes visible (Fig 5B), highlighting the limitations of unsupervised approaches in recognizing these boundaries independently. Another example is A5, where, without labeling, four distinct clusters are visible, suggesting the presence of four fibers/tracks in this recording (Fig 5C). However, when the spikes are labeled (Fig 5D), it becomes clear that only two fibers are in this recording, and they can be accurately separated using supervised methods with an accuracy of 0.99.

These results underscore the need for caution among microneurography researchers, as relying on unsupervised methods like PCA clustering may yield unreliable classifications in this context. We advocate alternative approaches or supervised methods that can leverage prior knowledge to improve classification accuracy when group boundaries are not inherently clear, as unsupervised methods may suggest more, or fewer classes than exist.

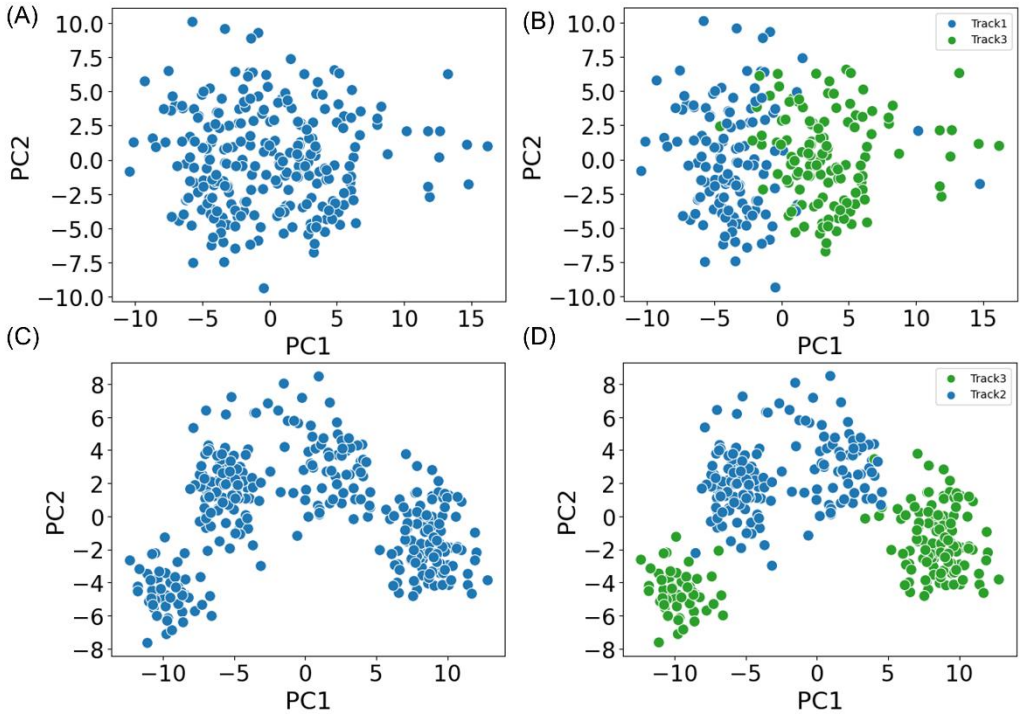


**Fig 5. PCA plots of recordings A1 (A and B) and A5 (C and D), illustrating the variability and challenges of clustering without prior labeling.** (A) Unlabeled data for recording A1, showing a single, indistinct cluster in the two-component PCA space with no clear separation among fiber/track classes. (B) The same data as in (A), now labeled by fiber/track, reveals underlying classes, illustrating that identifiable decision boundaries become visible only with prior knowledge of class labels. (C) Unlabeled data for recording A5, showing four clusters in the PCA feature space, suggesting the potential activity of four separate fibers/tracks. (D) The same data as in (C) with labeled spikes shows that these clusters correspond to only two actual classes, each with a well-defined separation. This separation allows for effective classification with a supervised method, achieving an accuracy of 0.99.

# 3 Discussion

Using human microneurography data from unmyelinated fibers, we demonstrate that supervised strategies for spike sorting, enabled by ground truth data, are a more reliable approach for spike sorting than unsupervised algorithms. Within the supervised approaches, the raw signal feature set ($W_{raw}$) revealed the greatest potential for sorting. It also outperformed feature sets derived from the SS-SPDF method, which have been successfully applied for spike sorting in prefrontal cortex recordings, suggesting effective methods for targeting the central nervous system may not be optimal for applications in the peripheral nervous system. Interestingly, simple morphological features also performed relatively well, which is likely a result of the microneurography-specific electrode placement strategy and maximizes visual differences in spike morphologies. Our findings emphasize that the choice of optimal spike sorting algorithm depends heavily on the characteristics of the specific electrophysiological method, as well as the specifics of individual recordings, which may exhibit significant variability. Importantly, the transparency of the computational pipeline allowed us to trace the relationships between the spike waveform similarities and the resulting sorting difficulties.

This work highlights the potential to develop “sortability” criteria for neurophysiological recordings based on the marking method employed. For instance, datasets in which the accuracy of tracked spikes falls below a predefined threshold could be declared as non-sortable and excluded from further analysis. The establishment of such thresholds needs to be tailored to the specific research objectives and the acceptable level of spike misclassification, ensuring alignment with the work's requirements.

Most spike sorting solutions are designed for extracellular recordings from the central nervous system, typically relying on feature extraction and clustering due to the common lack of ground truth data [17, 27]. Such recordings, with known ground truth, are rare as simultaneous intra- and extracellular recordings are required [28], so the validation of spike sorting algorithms is often limited to synthesized datasets. In the case of microneurography, the standard presence of baseline electrical stimulation can be leveraged to provide a sufficient amount of labeled data per recording. This makes spike sorting a classification task rather than clustering and allows the creation of benchmark datasets to accurately validate sorting methods. These electrically evoked spikes can be unambiguously sorted by their response latency, which reflects the time taken for the spikes to travel from the stimulation site to the recording electrode. Our findings demonstrate that clustering alone is not a reliable approach for analyzing microneurography data, indicating that this method cannot be fully trusted for accurate spike sorting. Unsupervised spike sorting algorithms are particularly unreliable for handling untracked spikes. Training a classifier with ground truth data could significantly enhance the reliability in sorting spikes that are not time-locked to their response latency.

The raw waveform ($W_{raw}$) yielded the best performance in 19 out of 26 datasets when employing SVM classifiers with RBF kernels. SVM was selected primarily for its minimal parameter requirements, which in turn reduces the amount of data needed for effective training. Moreover, SVMs have proven to be particularly effective in handling high-dimensional data, mainly when

applied to smaller, balanced datasets, making them an excellent fit for our recordings [29]. Notably, using a random forest classifier achieved higher accuracy in one feature set ($SPDF_{raw}$), indicating potential advantages in specific configurations. Full performance metrics for the random forest classifier are provided in the Supporting Information (files S5-S8) to allow for a detailed comparison. Surprisingly, the features from the SS-SPDF method ($SPDF_2$- $SPDF_{raw}$), which we expected to outperform other feature sets due to their wide range of spike characteristics, did not show superior performance in these microneurography data. This could be the result of morphological differences between central and peripheral neurons, the result of the high noise level in microneurography, or a suboptimal choice in a setup with a relatively small number of labeled spikes per recording.

While the accuracies achieved with $W_{raw}$ were significantly better than those obtained with all other feature sets, it is not optimal for every dataset. This indicates that the performance of a feature set can vary based on the specific characteristics of the data, highlighting that there is no universal solution for all recordings. While algorithmic approaches can be proposed for decision optimization, optimal sorting accuracy remains dependent on the recording conditions and fiber characteristics. The development of machine learning methods, such as automatic feature extraction [30] and transfer learning [31], could offer an effective strategy for combining multiple recordings into a unified model despite the observed sensitivity to feature set selection.

The performance of the SVM classifier significantly decreases as the number of fibers in a recording increases, which aligns with the documented issue in central nervous system recordings [32]. One potential approach to mitigate this challenge is a one-vs-all strategy, where only spikes from a single nerve fiber with a high signal-to-noise ratio are prioritized for sorting. However, some recordings may be too complex or noisy to allow for effective spike sorting, making them unsortable. It is, therefore, crucial to raise awareness among experimenters about the importance of minimizing the number of fibers in recordings when reliable spike sorting is essential. Prioritizing recordings with fewer fibers can help ensure more accurate and consistent spike sorting results.

Finally, this work is the first systematic spike morphology and spike sorting analysis of microneurography recordings. It stands out by harmonizing microneurography data from two geographically distant locations and independent work groups with different recording hardware and software [33]. We incorporated 26 datasets, which capture the typical broad spectrum of variability inherent to microneurography recordings. The dataset diversity was carefully curated to include a wide range of recording durations, fiber, and track counts, excluding recordings where fibers overlapped, as these instances could cause even more challenges for sorting. Spikes recorded via microneurography are highly sensitive to confounders, such as electrode movement, sudden overlay with other spikes originating from nerve fibers from the autonomic nervous system, or spontaneously discharging temperature-sensitive neurons and environmental electrical noise. Our analysis aimed to address these complexities while providing a transparent open-source pipeline, which can be adjusted to other types of neuro-electrophysiological data.

# 4 Conclusion and outlook

We conducted the first systematic investigation of spike sorting perspectives and challenges in extracellular single nerve fiber recordings from peripheral nerves. By harmonizing data between two microneurography labs and using an electrical stimulation protocol based on the marking method, we created a diverse ground-truth data collection. We designed an open-source computational pipeline to test various sorting approaches.

Our results show promising performance of the SVM algorithm applied to the raw spike signal in comparison to a wide variety of other extracted feature sets. Due to the high variance between datasets, we used individually trained models. The performance on each dataset largely depended on the number of neural fibers contributing to the single-channel signal and the morphological differences between spikes from different fibers.

Our approach is transferable to other neuro-electrophysiological recordings, particularly when regular stimulation that evokes single spikes can be employed to gain time-locked responses. This enables the generation of a reliable ground truth dataset for method validation.

In future work, we will explore the performance impact of testing combinations of all feature sets, moving beyond the distinct feature sets analyzed here to gain deeper insights into their contributions to model performance. We plan to benchmark our approach against existing spike sorting frameworks, such as Spike2, to evaluate their performance and suitability for microneurography data. This comparison will provide critical insights into standard pipelines' effectiveness and their adaptability to this recording method. However, a key limiting factor remains the challenge of spike detection in microneurography recordings. The high noise levels and low signal-to-noise ratio make many spikes small and undetectable using standard detection methods without the aid of the marking method. Addressing this issue is a priority in our ongoing work, as resolving it will be crucial to ensuring accurate comparisons and improving the overall reliability of microneurography data analysis. Further, we will explore deep-learning methods and fine-tuning in order to include multi-recording data in one model rather than retraining models for each dataset individually. Additionally, ADS-related parameters, which reflect the number and timing of previous fiber activity [15], could be used to enhance the model's performance probabilistically.

The knowledge gained from this work also allows an immediate step towards sorting and analyzing pain and itch-linked peripheral activity, which was not evoked by electrical stimulation. This can be achieved by setting high accuracy thresholds and only analyzing recordings with a satisfactory level of reliability. Furthermore, spike train features should be selected with consideration of their sensitivity to potential misclassifications. For good research practice, the information about the sorting accuracy of the time-locked spiked should be stored together with the spiking activity analysis as one of the assessments of the results’ reliability.

# Methods

## Microneurography

Twenty-six recordings from healthy volunteers were included in this work for the microneurography experiments. The studies involving human participants were reviewed and approved by the Ethics Board of the University Hospital RWTH Aachen with numbers Vo-Nr. EK141-19 and from the Faculty of Biomedical Sciences Research Ethics Committee at the University of Bristol (reference number: 51882). The participants provided their written informed consent, and the study was conducted according to the Declaration of Helsinki.

In both laboratories in Aachen and Bristol, a microelectrode (Frederick-Haer, Bowdoinham, ME, USA) is inserted into the superficial peroneal nerve (Fig 6A) while the volunteer or patient is awake and responsive. In Aachen, the signal is amplified and filtered with a Neuro Amp EX (ADInstruments) amplifier, an additional bandpass filter with 500 to 1000 Hz, and a 50 Hz notch filter. The recordings were acquired with 10 kHz with an analog digital converter from National Instruments and a customized software Dapsys (www.dapsys.net) from Brian Turnquist [5, 34]. In the laboratory in Bristol, a custom-made recording system and software were used based on electronics and software from Open Ephys [35] and APTrack [36]. This system acquires at 30 kHz, and the signal is digitally filtered (300 to 6000 Hz). The acquisition board was electrically isolated from the system using a 5kV optoisolator (Intona, Germany). In both laboratories, the receptive field of C-fibers is determined through transcutaneous electrical stimulation utilizing a Digitimer DS7 constant current stimulator. Once the receptive field is identified, C-fibers are repetitively stimulated at a low frequency (e.g., 0.25 Hz), which enables the marking method to be used. A more detailed description of microneurography can be found here [37].

In this work, spikes were elicited through electrical stimuli. We recorded 22 datasets in the microneurography laboratory in Aachen (A1-A22), and four files were acquired at the lab in Bristol (B1-B4). Details are listed in Table 2. These datasets have different recording conditions, including varying noise levels, signal complexity, and the number of fibers/tracks. Each dataset presents distinct challenges, for example, spikes from different tracks have the same amplitude or there are more than two fibers/tracks in a single dataset. This diverse collection provided a comprehensive set for evaluating the sorting results.

**Table 2** Datasets overview, including the number of active fibers/tracks, the number of spikes, and the class distributions. The classes are mostly equally distributed. Datasets labeled with A are from the lab in Aachen and datasets labeled with B are from Bristol.

| Dataset | Number of fibers/tracks | Number of spikes | Fiber/track labels and distribution |
|---|---|---|---|
| A1 | 2 | 258 | Track1: 129<br>Track3: 129 |
| A2 | 2 | 536 | Track3: 268<br>Track4: 268 |
| A3 | 2 | 346 | Track1: 173<br>Track2: 173 |

| | | | |
|---|---|---|---|
| A4 | 3 | 1966 | Track1: 647<br>Track2: 632Track4: 687 |
| A5 | 2 | 340 | Track2: 170<br>Track3: 170 |
| A6 | 2 | 206 | Track2: 103<br>Track4: 103 |
| A7 | 3 | 258 | Track2: 86 Track4: 86<br>Track6: 86 |
| A8 | 3 | 300 | Track3: 102 Track4: 97<br>Track7: 101 |
| A9 | 2 | 230 | Track1: 115<br>Track2: 115 |
| A10 | 3 | 419 | Track2: 152<br>Track4: 153 Track5: 114 |
| A11 | 5 | 465 | Track3: 93 Track4: 93<br>Track13: 93 Track15: 93<br>Track22: 93 |
| A12 | 6 | 1932 | Track3: 348 Track7: 350<br>Track8: 350 Track11: 349<br>Track13: 344 Track15: 348 |
| A13 | 4 | 508 | Track1: 124 Track5: 124<br>Track11: 129 Track20: 128 |
| A14 | 2 | 310 | Track1: 155<br>Track2: 155 |
| A15 | 2 | 283 | Track2: 137<br>Track7: 146 |
| A16 | 5 | 800 | Track1: 159 Track3: 159<br>Track5: 160 Track7: 157<br>Track8: 160 |
| A17 | 5 | 963 | Track1: 205 Track2: 152<br>Track3: 205 Track4: 203<br>Track6: 198 |
| A18 | 5 | 1036 | Track1: 209 Track2: 208<br>Track3: 209 Track4: 206<br>Track6: 204 |
| A19 | 4 | 816 | Track1: 205 Track3: 205<br>Track4: 204 Track7: 202 |
| A20 | 3 | 342 | Track1: 122<br>Track4: 108 |
| A21 | 2 | 526 | Track3: 218<br>Track4: 308 |
| A22 | 2 | 288 | Track3: 144<br>Track4: 144 |
| B1 | 3 | 1082 | R: 277 S: 499<br>T: 306 |
| B2 | 5 | 981 | unit_0: 252 unit_1: 254<br>unit_2: 232 unit_5: 243 |
| B3 | 3 | 513 | unit_0 189 unit_1 183<br>unit_2 141 |
| B4 | 2 | 366 | unit_0 184<br>unit_1 182 |

Table 2 provides a summary of the key characteristics of each dataset, including the number of tracked fibers and the total spike count. The spike count per recording ranges from 206 to 1966, representing the data points that will be split into training and test sets. The number of fibers/tracks

varies from two up to six, reflecting the complexity of each dataset. The final column shows the class distribution, indicating that the classes are generally balanced across most recordings.

## Marking Method

In microneurography, experimenters employ the marking method [15] to observe nerve fiber responses in the form of spikes. The technique utilizes the characteristic that C-fibers have an almost constant conduction velocity in response to repetitive low-frequency stimulation (e.g., 0.125-0.25 Hz), described here as background stimulation. This allows time-locking a subset of spikes. When raw signal segments are plotted sequentially and vertically in a waterfall representation, where each segment starts at the onset of the background stimulus, the spike responses evoked by the background stimulation are vertically aligned (red and green spikes in Fig 6B). This alignment enhances the visibility of spikes, even when the signal-to-noise ratio is poor (Fig 6C for an exemplary spike).

Different C-fibers show distinctive conduction velocities, facilitating the differentiation of multiple fibers within a single recording [5]. When applying further stimulation in the form of extra electrical pulses or natural stimuli, there is a slowing in the conduction velocity of the signal transmission and an increase in latency to the subsequent background stimulus. This phenomenon is known as activity-depending slowing (ADS) [16] and “marks” the fiber. ADS is useful not only to distinguish fibers but also for their classification as different physiological classes of C-fiber exhibit differing degrees of slowing to low and high-frequency electrical stimulation, such as mechanosensitive (CM) or mechanoinsensitive (CMi) C-fibers [37]. In Fig 6B, a representative “waterfall” plot is presented, illustrating the trajectories of two tracked fibers. In this manuscript, we call them *tracks*. When the stimulation remains constant (indicated by blue rectangles), the responses align vertically (lines 1-5). However, after stimulating the fibers with two extra pulses (as seen in line 6), we can observe ADS in both fibers. Through subsequent repetitive low frequency stimulation, latencies recover to their initial values. The remaining issues and challenges in accurately sorting spontaneous firing or chemically induced activity when the spikes have a similar shape persist if in more than one fiber ADS is observed (see line 11). The red and green spikes, representing responses to extra stimuli in line 6, and the orange-marked spontaneous activity in line 11, cannot be reliably sorted as they are not on the tracks.

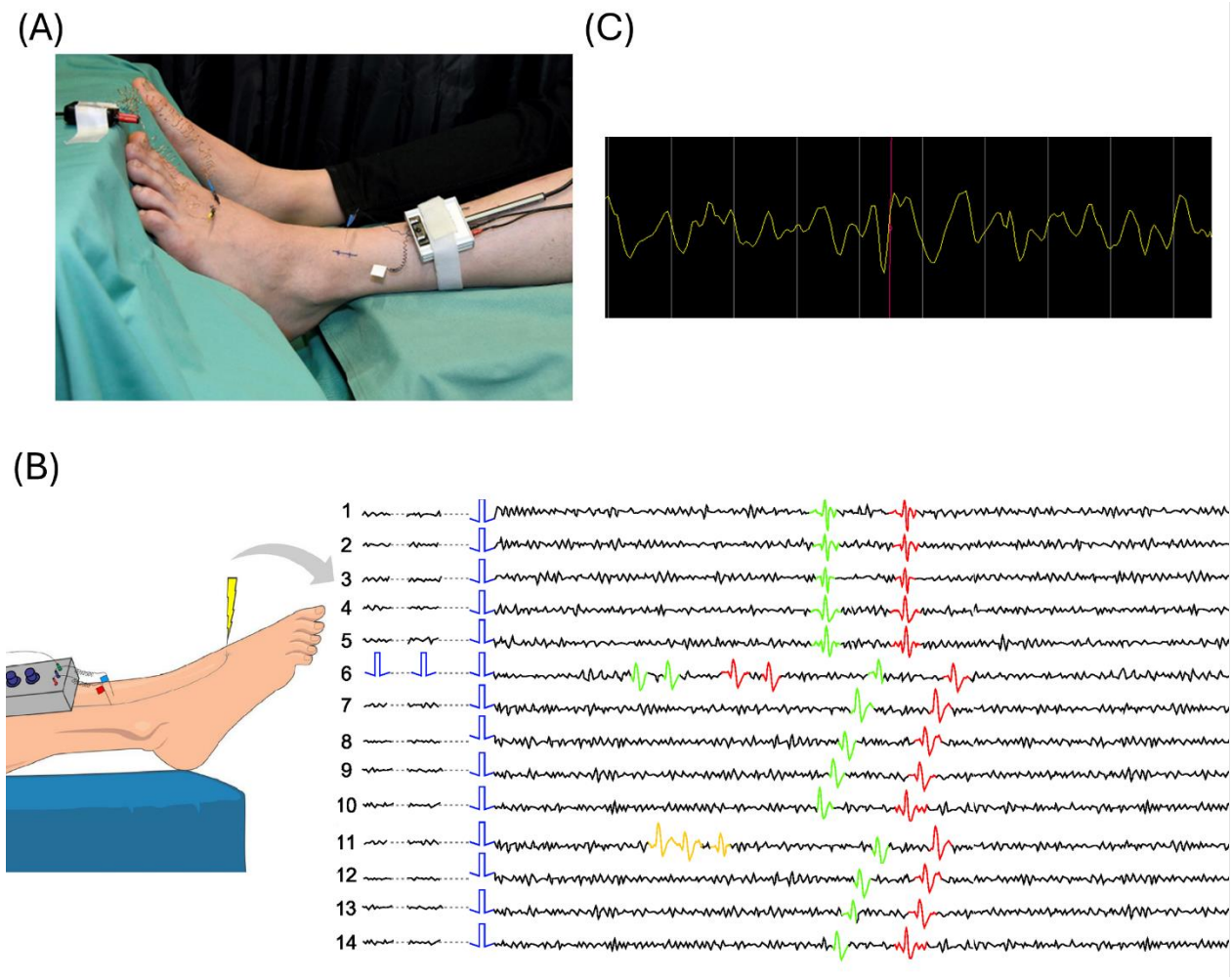


**Fig 6. Details on microneurography experiments.** (A) Set up of microneurography experiments. One microelectrode is inserted into a fascicle of the nerve and another as a reference electrode in the skin nearby. In the receptive field in the skin, C-fibers are activated by electrical stimulation. (B) Schematic waterfall plot of two C-fiber tracks. The waterfall provides a visual representation of the marking method with two active fibers (green and red spikes) represented as track. The onset of the electrical stimulation is indicated by the blue rectangle. Each line, referred to as a trace, begins with the low-frequency stimulus. When extra stimulation is applied (line 6, red and green spikes) or spontaneous activity occurs (line 11, orange spikes), ADS is observed in both fibers [38]. (C) An example spike with a low signal-to-noise ratio. The screenshot of a recording in Dapsys, in which the red line indicates the exemplary spike of interest with an amplitude similar to the background noise level. We could only identify the spike by the marking method.

## Pre-processing for feature extraction

To ensure data harmonization, we agreed on utilizing the HDF5/NIX [26, 39] format, a well-established standard in the field of electrophysiology.

To handle the datasets generated by Dapsys, we used our Python package PyDapsys [23]. This package enables us free access to electrophysiological recordings stored in Dapsys' proprietary data format. For analyzing an individual recording, we retrieved the raw signal with timestamps and voltages, the onset timestamps of all tracked spikes with their corresponding track label to ensure

ground truth data, as well as the onset timestamps of stimulation events. The data were processed in Panda's data frames.

To facilitate the extraction of spike feature sets, our next step involved obtaining the tracked spike waveforms from the raw signal. As illustrated in Fig 6C, the red line represents the reference point in time, which is typically located near the center of the waveform. Considering an extracellular C-fiber spike width of approximately 3 ms and a sampling frequency of 10,000 Hz for Dapsys files, it is necessary to encompass 30 datapoints from the raw signal to adequately capture the spike. For instance, a spike occurring at timestamp t would lie within the data slice window [t – 15, t + 15]. Conversely, for Bristol files with a sampling frequency of 30,000 Hz, we consider 90 datapoints, thereby expanding the window range to [t – 30, t + 60], as the reference point is slightly shifted to the left. As each spike is identified by a timestamp, we extracted the corresponding index, which we consider to be the center of the spike in the raw signal. This index was then used to determine the voltage values of the spike with the 3 ms window. To ensure the correct alignment of all spikes, we established that this center point index is two positions before the voltage maximum of the spike, which should occur in the rising phase of the spike.

To gain deeper insights into each track's characteristic shape, we compute a “template” by averaging all the individual spikes associated with that specific track. This averaging process results in a representative waveform that effectively shows the tracks' distinctiveness or similarity in shape. The templates from two exemplary recordings are shown in Fig 4.

We have implemented automatic outlier removal for the features from the SS-SPDF method by comparing each spike to its corresponding template. A spike is removed if its minimum value falls outside a ±30% range of the template's minimum value. The full pre-processing steps can be found in Table 3.

**Table 3** Pre-processing pipeline for Aachen and Bristol data. For the Aachen sampling frequency (10 kHz), 30 datapoints correspond to 3 ms in time. For the Bristol frequency (30 kHz), 90 datapoints correspond to 3 ms.

| **Pre-processing Aachen** | **Pre-processing Bristol** |
|---|---|
| Create NIX file from Dapsys file | Create NIX file from raw Open Ephys file or use H5 file |
| Read-in signal from raw Dapsys file, read-in stimuli and tracks from NIX file | Read-in signal, stimuli, and tracks from NIX file |
| | Smooth signal (running mean with window size 4) and invert signal (Bristol signal is flipped) |
| Create data frame with stimuli | Create data frame with stimuli |
| Get signal values for each AP based on timestamp ***t*** in a window ***[t - 15, t + 15]*** | Get signal values for each AP based on timestamp ***t*** in a window ***[t - 30, t + 60]*** |
| Find maximum voltage in signal piece $m$ to align spikes by index $m - 2$ | Find maximum voltage in signal piece $m$ to align spikes by index $m - 2$ |

| | |
|---|---|
| Increase sampling frequency by resampling 30 datapoints to 60 | |
| Compute template and determine threshold for each template for automatic filtering of spikes | Compute template and determine threshold for each template for automatic filtering of spikes |
| Compute components for complex features by taking the max value of the first derivative between 20:40 position | Compute components for complex features by taking the max value of the first derivative after 30th data point |
| Compute all feature sets | Compute all feature sets |
| Plot AP shapes with template | Plot AP shapes with template |

## Feature set extraction

In our analysis, we explored various feature extraction techniques aimed at quantifying the characteristics of spike waveforms (Table 4). We compared the disparities between simpler and more sophisticated feature sets. These extracted feature sets served as inputs for classification methods, and we evaluated their effectiveness in characterizing and discriminating spikes from multiple tracks.

**Table 4** Different feature sets are extracted through domain-agnostic dimension reduction as well as through spike approaches, using the raw waveform and computed features on the waveforms. The feature sets include methodologies, such as principal component analysis (PCA) and the features from the spike sorting approach based on shape, phase, and distribution features (SS-SPDF) by Caro-Martín et al. [19].

| Feature Set | Description | Number of features |
|---|---|---|
| Simple | Amplitude and width | 2 |
| $SPDF_2$ | PCA of SS-SPDF features (2-comp) | 2 |
| $SPDF_3$ | PCA of SS-SPDF features (3-comp) | 3 |
| $SPDF_{raw}$ | Raw SS-SPDF features | 23 |
| $W_2$ | PCA of raw waveform features (2-comp) | 2 |
| $W_3$ | PCA of raw waveform features (3-comp) | 3 |
| $W_{raw}$ | Raw waveform features | 30/90 |

### Amplitude and width – Simple

We refer to simple features when taking the most fundamental characteristics of a waveform. Here, we examined two specific features: amplitude $a$ and width $w$. Amplitude is defined as the positive peak voltage value of a spike, while the spike width is defined as the distance between the two nearest points where the signal falls below half of the spike's amplitude. Consequently, each spike can be described by a two-dimensional feature vector. Formally, the feature vector for a spike $S$ is denoted as:

$$S = [a, w].$$

**Shape-, phase- and distribution-based features – $SPDF_2$- $SPDF_{raw}$**

In 2018, Caro-Martín et al. developed a comprehensive spike sorting pipeline for extracellular recordings that outperforms contemporary methods used in neurophysiology [19]. They tested their algorithm on two synthetic datasets as well as on real extracellular recordings of neural activity within the rostral-medial prefrontal cortex from rabbits. Within their methods, they extracted 24 distinct features from each spike on shape-, phase-, and distribution-based characteristics. Shape features describe the waveform of the shapes's first derivative in the time domain, while phase features relate the amplitudes of the first and the second derivative to points in the phase space. Additionally, distribution features take into account the amplitude distributions of both the first and second derivatives of each spike.

Caro-Martín et al. employed a modified k-means clustering algorithm that finds the optimal number of clusters and clustering results. This algorithm also addresses the issue of overlapping spikes and evolving waveforms over time. They introduced their own validity score and error index to evaluate their method and compare it with alternative spike sorting methods. Furthermore, the authors emphasize the method's computational efficiency and highlight the enhanced physiological interpretability of their feature-based approach compared to algorithms relying on dimensionality reduction techniques.

In this work, we implemented these features in Python. They represent the more sophisticated approach. In our implementation, we primarily followed the definitions in the paper. However, we had to make certain necessary adjustments.

First, we compute both the first and second derivatives for each spike. Six fundamental points characterize a spike, forming the base for the feature computations. In some of our detected instances, the first fundamental point (the first zero-crossing of the first derivative) was undefined. Consequently, we automatically excluded these spikes from our data. Furthermore, to enhance the precision of the derivative computations, we resampled the spikes originally recorded with Dapsys from 30 to 60 datapoints with SciPy's [40] resample function.

For the final implementation, we had to modify two feature definitions. In the case of Feature 4, it was unclear how the reference waveform was computed. Therefore, we decided to exclude Feature 4 from our final vector.

For Feature 8, the original description referred to it as the "root-mean-square of the amplitudes before the FD event of the action potential" [19]. However, after closer examination of the mathematical formula, it became apparent that the values were not squared, and the authors did not define the variable $m$. We redefined Feature 8 as $f_8 = \sqrt{\frac{\sum_{i=s}^{P1} a_{FD_i}^2}{P1-s}}$, where $s$ is the beginning of the window.

The final feature vector of spike $S$ is denoted as follows:

$$S = [f_1, f_2, f_3, f_5, \dots, f_{24}].$$

### Raw signal – $W_2$ - $W_{raw}$

The next feature set is the raw signal itself. We did not process the signal segments further and, instead, directly employ the 30 and 90 voltage values, respectively, as input for each spike waveform. As a result, for a given spike denoted as *S*, the feature vector is denoted as follows with $s$ describing the voltage at position *x*:

$$S_{Aachen} = [s_1, s_2, \dots, s_{30}],$$

$$S_{Bristol} = [s_1, s_2, \dots, s_{90}].$$

### PCA features of raw signal and complex feature sets

Due to the discrepancy in the dimensionalities of the initially presented feature set vectors, we employed principal component analysis (PCA) on the raw signal and the feature vector defined by Caro-Martín et al. We conducted PCA with both two and three components to reduce the feature vector from its original 23, 30, or 90 dimensions, respectively. As a result, we generated four additional feature vectors for each spike, all of which serve as input for the spike sorting process. The definitions of these feature vectors for the spike are denoted as follows:

$$S_{2\ components} = [PC1, PC2] \text{ and } S_{3\ components} = [PC1, PC2, PC3].$$

## Classification

As a classification method, we applied support vector machines (SVMs) motivated by their efficiency and low number of hyperparameters [41]. The models aim to identify a line or hyperplane within the feature space to distinguish between different classes. To assess the accuracy and generalizability of our models, we employed 5-fold cross-validation, which repeatedly partitions the data into training and test sets. We adhered to the default parameters as provided in the sci-kit learn implementation [42].

### Accuracy

To assess our spike sorting results and compare the performance of various feature sets and methodologies, we used evaluation methods tailored to each approach. Accuracy is a metric for evaluating classification results and describes the fraction of correct predictions [43]. We employed accuracy as it is most meaningful when class sizes are relatively balanced. Higher accuracy indicates a better-performing classification model. It is usually expressed as a percentage and is defined as the ratio of correctly predicted instances to the total instances:

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions} \qquad (1)$$

## Statistical analysis

To evaluate whether there were statistically significant differences between the accuracies of the feature sets, we employed the Wilcoxon signed-rank test, a non-parametric test used for comparing two matched samples [44]. This test was selected because it does not assume a normal distribution of the data and is appropriate for comparing paired observations, such as accuracy scores obtained from different feature sets on the same datasets. The statistical analysis was performed with SciPy [40] v1.10.1.

The accuracies for each feature set were compared in a pairwise manner with each pair representing the performance of two feature sets on the same recordings. The test was used to determine whether the distribution of the accuracy differences between the feature sets was statistically significant.

Given that multiple pairwise comparisons were performed, the risk of false positive errors increases. To address this, we applied the Bonferroni correction to adjust the alpha level [45]. Specifically, the initial alpha level ($\alpha = 0.01$) was divided by the number of comparisons to control the error rate ($\alpha_{adj} = 0.01/n$, where $n$ is the number of comparisons). The corrected alpha level was used to determine whether the differences between the feature sets were statistically significant after adjustment for multiple comparisons.

## Measurement of template similarity

To assess the similarity between spike templates, we used several distance metrics to quantify the differences between two spike templates. These templates are the average waveform of spikes associated with individual fibers.

The motivation for using these distance metrics came from the need to investigate whether the similarity between spike templates correlates with the classification accuracy of the feature sets derived from these waveforms. Specifically, we hypothesized that feature sets generated from more similar spike templates would yield lower classification accuracies, while feature sets generated from more distinct templates would lead to higher accuracies.

We compared the distance metrics between spike templates with the best classification accuracies to test this hypothesis. By evaluating the relationship between the template similarities, we aimed to determine whether lower similarity between spike templates predicts better classification performance.

We considered several distance metrics in this analysis, including mean squared error (MSE) [46], mean absolute error (MAE) [47], and root mean squared error (RMSE) [48]. Here, $n$ represents the length of each template. $T = [t_1, \dots, t_n]$ and $\hat{T} = [\hat{t}_1, \dots, \hat{t}_n]$ denote the waveform templates being compared.

### Mean Squared Error

The mean squared error (MSE) is used to quantify the average squared difference between two waveform templates. Smaller MSE values indicate higher similarity, and larger values denote greater distinguishability between templates. For this analysis, MSE was selected as it effectively highlights differences in waveform shape.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(t_i - \hat{t}_i)^2 \qquad (2)$$

### Mean Absolute Error

The mean absolute error (MAE), on the other hand, measures the average absolute difference between templates, offering a straightforward interpretation of the average error without squaring the values.

$$MAE = \frac{\sum_{i=1}^{n}|t_i - \hat{t}_i|}{n} \qquad (3)$$

### Root Mean Squared Error

The root mean squared error (RMSE) provides a distance measure in the same units as the original data, making it easier to interpret in practical terms.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(t_i - \hat{t}_i)^2}{n}} \qquad (4)$$

## Acknowledgements

This project was supported by a grant from the Interdisciplinary Center for Clinical Research within the faculty of Medicine at the RWTH Aachen University. We acknowledge Aidan Nickerson and Danxia Bao for their technical support.

## References

[1] Ackerley R, Watkins RH. Microneurography as a tool to study the function of individual C-fiber afferents in humans: responses from nociceptors, thermoreceptors, and mechanoreceptors. J Neurophysiol. 2018 Dec 1; 120(6): 2834–46.

[2] Wei Y, Marshall AG, McGlone FP, Makdani A, Zhu Y, Yan L, et al. Human tactile sensing and sensorimotor mechanism: from afferent tactile signals to efferent motor control. Nat Commun. 2024 Aug 10; 15(1): 6857.

[3] Yger P, Spampinato GL, Esposito E, Lefebvre B, Deny S, Gardella C, et al. A spike sorting toolbox for up to thousands of electrodes validated with ground truth recordings in vitro and in vivo. eLife. 7: e34518.

[4] Horváth C, Tóth LF, Ulbert I, Fiáth R. Dataset of cortical activity recorded with high spatial resolution from anesthetized rats. Sci Data. 2021 Jul 15; 8(1): 180.

[5] Turnquist B, RichardWebster B, Namer B. Automated detection of latency tracks in microneurography recordings using track correlation. J Neurosci Methods. 2016 Mar 15; 262: 133–41.

[6] Buccino AP, Hurwitz CL, Garcia S, Magland J, Siegle JH, Hurwitz R, et al. SpikeInterface, a unified framework for spike sorting. Colgin LL, Grün S, Kloosterman F, editors. eLife. 2020 Nov 10; 9: e61834.

[7] Harris KD, Quian Quiroga R, Freeman J, Smith S. Improving data quality in neuronal population recordings. Nat Neurosci. 2016 Aug 26; 19(9): 1165–74.

[8] Krahe R, Gabbiani F. Burst firing in sensory systems. Nat Rev Neurosci. 2004 Jan; 5(1): 13–23.

[9] Torebjörk E. Nociceptor activation and pain. Philos Trans R Soc Lond B Biol Sci. 1985 Feb 19; 308(1136): 227–34.

[10] Torebjörk HE, Hallin RG. Responses in human A and C fibres to repeated electrical intradermal stimulation. J Neurol Neurosurg Psychiatry. 1974 Jun; 37(6): 653–64.

[11] Vallbo AB, Hagbarth KE. Activity from skin mechanoreceptors recorded percutaneously in awake human subjects. Exp Neurol. 1968 Jul; 21(3): 270–89.

[12] Kutafina E, Becker S, Namer B. Measuring pain and nociception: Through the glasses of a computational scientist. Transdisciplinary overview of methods. Front Netw Physiol. 2023; 3.

[13] Kleggetveit IP, Namer B, Schmidt R, Helås T, Rückel M, Ørstavik K, et al. High spontaneous activity of C-nociceptors in painful polyneuropathy. Pain. 2012 Oct; 153(10): 2040–7.

[14] Forster C, Handwerker HO. Automatic classification and analysis of microneurographic spike data using a PC/AT. Journal of Neuroscience Methods. 1990 Feb 1; 31(2): 109–18.

[15] Schmelz M, Forster C, Schmidt R, Ringkamp M, Handwerker HO, Torebjörk HE. Delayed responses to electrical stimuli reflect C-fiber responsiveness in human microneurography. Exp Brain Res. 1995; 104(2): 331–6.

[16] Serra J, Campero M, Ochoa J, Bostock H. Activity-dependent slowing of conduction differentiates functional subtypes of C fibres innervating human skin. J Physiol. 1999 Mar 15; 515(Pt 3): 799–811.

[17] Gibson S, Judy JW, Marković D. Spike Sorting: The First Step in Decoding the Brain: The first step in decoding the brain. IEEE Signal Process Mag. 2012 Jan; 29(1): 124–43.

[18] Adamos DA, Kosmidis EK, Theophilidis G. Performance evaluation of PCA-based spike sorting algorithms. Comput Methods Programs Biomed. 2008 Sep 1; 91(3): 232–44.

[19] Caro-Martín CR, Delgado-García JM, Gruart A, Sánchez-Campusano R. Spike sorting based on shape, phase, and distribution features, and K-TOPS clustering with validity and error indices. Sci Rep. 2018 Dec 12; 8(1): 17796.

[20] Troglio A, Nickerson A, Schlebusch F, Röhrig R, Dunham J, Namer B, et al. odML-Tables as a Metadata Standard in Microneurography. Stud Health Technol Inform. 2023 Sep 12; 307: 3–11.

[21] Grewe J, Wachtler T, Benda J. A Bottom-up Approach to Data Annotation in Neurophysiology. Front Neuroinform. 2011; 5. Available from: https://www.frontiersin.org/articles/10.3389/fninf.2011.00016

[22] Sprenger J, Zehl L, Pick J, Sonntag M, Grewe J, Wachtler T, et al. odMLtables: A User-Friendly Approach for Managing Metadata of Neurophysiological Experiments. Front Neuroinform.; 13. Available from: https://www.frontiersin.org/articles/10.3389/fninf.2019.00062

[23] Konradi P, Troglio A, Pérez Garriga A, Pérez Martín A, Röhrig R, Namer B, et al. PyDapsys: an open-source library for accessing electrophysiology data recorded with DAPSYS. Front Neuroinform. 2023; 17. Available from: https://www.frontiersin.org/articles/10.3389/fninf.2023.1250260

[24] Schlebusch F, Kehrein F, Röhrig R, Namer B, Kutafina E. openMNGlab: Data Analysis Framework for Microneurography - A Technical Report. Stud Health Technol Inform. 2021 Sep 21; 283: 165–71.

[25] Troglio A, Kutafina E, Namer B. SpikeSortingPipeline. Available from: https://github.com/Digital-C-Fiber/SpikeSortingPipeline

[26] Adrian S, Kellner C, Jan B, Thomas W, Grewe J. File format and library for neuroscience data and metadata. Front Neuroinform. 2014 Jan 1; 8.

[27] Rey HG, Pedreira C, Quian Quiroga R. Past, present and future of spike sorting techniques. Brain Res Bull. 2015 Oct 1; 119: 106–17.

[28] Einevoll GT, Franke F, Hagen E, Pouzat C, Harris KD. Towards reliable spike-train recordings from thousands of neurons with multielectrodes. Curr Opin Neurobiol. 2012 Feb 1; 22(1): 11–7.

[29] Huang FJ, LeCun Y. Large-scale Learning with SVM and Convolutional for Generic Object Categorization. IEEE Comput Soc Conf Comput Vis Pattern Recognit. 2006. 284–91. Available from: https://ieeexplore.ieee.org/document/1640771

[30] Feature Extraction By Using Deep Learning: A Survey. IEEE Xplore. Available from: https://ieeexplore.ieee.org/abstract/document/8474912

[31] Weiss K, Khoshgoftaar TM, Wang D. A survey of transfer learning. J Big Data. 2016 May 28; 3(1): 9.

[32] Pedreira C, Martinez J, Ison MJ, Quian Quiroga R. How many neurons can we see with current spike sorting algorithms? J Neurosci Methods. 2012 Oct 15; 211(1): 58–65.

[33] Villa S, Aasvang EK, Attal N, Baron R, Bourinet E, Calvo M, et al. Harmonizing neuropathic pain research: outcomes of the London consensus meeting on peripheral tissue studies. Pain. 2024 Oct 16; Available from: 10.1097/j.pain.0000000000003445.

[34] Turnquist B. DAPSYS (Data Acquisition Processor System). Available from: http://dapsys.net/

[35] Siegle JH, López AC, Patel YA, Abramov K, Ohayon S, Voigts J. Open Ephys: an open-source, plugin-based platform for multichannel electrophysiology. J Neural Eng. 2017 Aug; 14(4): 045003.

[36] Nickerson AP, Newton GWT, O'Sullivan JH, Martinez-Perez M, Sales AC, Williams G, et al. Open-Source Real-Time Closed-Loop Electrical Threshold Tracking for Translational Pain Research. J Vis Exp. 2023 Apr 21; 194.

[37] Fiebig A, Leibl V, Oostendorf D, Lukaschek S, Frömbgen J, Masoudi M, et al. Peripheral signaling pathways contributing to non-histaminergic itch in humans. Journal of Translational Medicine. 2023 Dec 12; 21(1): 908.

[38] Kutafina E, Troglio A, de Col R, Röhrig R, Rossmanith P, Namer B. Decoding Neuropathic Pain: Can We Predict Fluctuations of Propagation Speed in Stimulated Peripheral Nerve? Front Comput Neurosci. 2022; 16. Available from: https://www.frontiersin.org/articles/10.3389/fncom.2022.899584

[39] Fortner B. HDF: The hierarchical data format. J Software Tools Prof Program; 1998.

[40] Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods. 2020 Mar; 17(3): 261–72.

[41] Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995 Sep 1; 20(3): 273–97.

[42] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. arXiv; 2018. Available from: http://arxiv.org/abs/1201.0490

[43] Opitz J. A Closer Look at Classification Evaluation Metrics and a Critical Reflection of Common Evaluation Practice. Trans Assoc Comput Linguist. 2024 Jun 25; 12: 820–36.

[44] Wilcoxon F. Individual Comparisons by Ranking Methods. Biometrics Bulletin. 1945;1(6):80–3.

[45] Weisstein EW. Bonferroni Correction [Internet]. Wolfram Research, Inc.; Available from: https://mathworld.wolfram.com/BonferroniCorrection.html

[46] Bickel PJ, Doksum KA. Mathematical Statistics: Basic Ideas and Selected Topics, Volumes I-II Package. New York: Chapman and Hall/CRC; 2015. 1066.

[47] Willmott CJ, Matsuura K. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate Research. 2005 Dec 19; 30(1): 79–82.

[48] Hyndman RJ, Koehler AB. Another look at measures of forecast accuracy. International Journal of Forecasting. 2006 Oct 1; 22(4): 679–88.

## Supporting Information

**S1 File. SVM classification accuracy scores.** Accuracy values for SVM classification, provided as a CSV file with comma-separated values.

**S2 File. SVM classification macro-averaged F1-scores.** Macro-averaged F1-scores values for SVM classification, provided as a CSV file with comma-separated values.

**S3 File. SVM classification precision scores.** Precision values for SVM classification, provided as a CSV file with comma-separated values.

**S4 File. SVM classification recall scores.** Recall values for SVM classification, provided as a CSV file with comma-separated values.

**S5 File. Random Forest classification accuracy scores.** Accuracy values for Random Forest classification, provided as a CSV file with comma-separated values.

**S6 File. Random Forest classification macro-averaged F1-scores.** Macro-averaged F1-scores values for Random Forest classification, provided as a CSV file with comma-separated values.

**S7 File. Random Forest classification precision scores.** Precision values for Random Forest classification, provided as a CSV file with comma-separated values.

**S8 File. Random Forest classification recall scores.** Recall values for Random Forest classification, provided as a CSV file with comma-separated values.

**S9 File. Computed templates and best SVM classification accuracy.** Computed templates for all recordings/datasets with two fibers, along with the maximum achieved accuracy for SVM classification. These templates were used to investigate how template similarity affects sorting quality.