**A study of non-linear feature extraction in spike sorting**

Eugen-Richard Ardelean1, Raluca Laura Portase1

1Department of Computer Science, Technical University of Cluj-Napoca, Cluj-Napoca, Romania

**\* Corresponding authors:** [ardeleaneugenrichard@gmail.com](mailto:ardeleaneugenrichard@gmail.com), [placehoder-raluca]

**ORCID Author IDs:**

Eugen-Richard Ardelean: 0000-0002-0098-4228

Raluca Laura Portase: 0000-0002-8985-4728

**Abstract: [**TBWritten]

**Keywords:** clustering, [TBAdded]

# Introduction

## Spike Sorting [TBRewritten]

Extracellular recordings capture the neural activity as voltage fluctuations from multiple nearby neurons (1), producing a continuous signal. Each individual activity of a single neuron is called a spike and in the case of extracellular recordings the neuron that generated such an activity is unknown. Spike sorting is the process of assigning each detected spike waveform (2) from an extracellular recording to its source neuron, based on the assumption that each neuron produces spikes of consistent shape (3), while different neurons generate distinguishable shapes from each other (2). However, the shape of spikes can be distorted by noise, electrode drift (4) and biological variability which result in scattered clusters instead of pinpoints. Consequently, it is important to find a robust approach to reduce the impact of these phenomena. One approach is to extract a small set of discriminative features that preserve the information that differentiates between the shapes of spikes to improve clustering.

The cannonical approach to spike sorting (2) is a pipeline of four sequantials steps: filtering, spike detection and clustering. Filtering is done in a band-pass manner and is applied to the raw recorded signal to isolate the 300 and 3000Hz (5) frequency band where spiking occurs. Spikes are detected typically through a simple amplitude thresholding based on the standard deviation of the signal multiplied by a scalar value (typically, from 3 to 5) to identify the events that could represent the neuronal activity. The choice of this threshold is a compromise between identifying noise segments as spikes and losing spikes of lower amplitude. Feature extraction is utilised to both generate a more auspicious space and to improve efficiency for clustering. The newly generated feature space should be robust as in invariant to small changes in shape such that it provides an easily separable space for clustering while simultaneously it should reduce the dimensionality to improve the execution time of clustering. As the final step of the spike sorting pipeline, the spikes should be grouped to represent all instances of activity of each individual neuron. This final assignment of spikes by source neurons is done by the clustering algorithm.

The spike sorting pipeline has seen many forms over the years. Initially, a supervised manual approach was taken where an expert would classify spikes based on a visual interpretation of similarity in a low dimensional space. This reduced space was generated by simple features (22) (26) such as amplitude, width, the peak-to-trough ratio. Using the peak-to-trough ratio feature was found to be useful in determining the type of neuron as narrow spikes (small peak-to-trough ratio) are representative of inhibitory neurons, while excitatory neurons have wider spikes (27). Since the 1950s, the number of recorded neurons has increased exponentially (6) rendering manual approaches unfeasible and recent developments in recording hardware (4,7) are following this trend. Through empirical analysis, probabilistics models were created that were able to leverage the entire spike waveform (23) allowing for the processing of a low number of electrodes. Later, the high-dimensional space of the spike waveform was projected to lower-dimensional spaces by applying PCA (24) and time-frequency transforms (such as the Wavelet Transform (25)) have started being used to introduce the frequency information in the computation.

The choice of approach is also depedent upon the nature of the analysis, offline or online. Offline spike sorting allows for the use of more complex algorithms as there is no time constraint as the analysis is done after the recording has finished. However, in an online setting, the algorithms employed must have to ability to process the data during the recording, thus requiring low execution times.

A common approach taken lately is template matching (8) on subsets of data, usually this approach substitutes the steps of spike detetion, feature extraction and even clustering in the canonical spike sorting pipeline. Due to it being applied on subsets rather than the whole data, it is an efficient approach from a computational perspective. One such algorithm that employs template matching is M-Sorter (28). M-Sorter is an automated approach to spike detection and classification based on coefficients obtained through the wavelet transform and template matching. This method can be seen as a two-stage process when applied to the filtered signal. It employs correlation of the wavelet coefficients for the detection of the spikes, while through the use of K-Means templates are generated which are used in the assignment of spikes to the neurons that produced them (considered to be that which has the smallest distance).

In this work, we attempt to examine the impact of feature extraction in spike sorting. Although, it is the clustering that outputs the final result and separation of the space into clusters, it is actually the feature extraction which must obtain a separable space for the clustering. Similarly to clustering algorithms, a golden standard (9,10) does not exist for feature extraction algorithms (2,5) either. Their performance depends on the particular set of characteristics of the input data. Here, we employ a number of non-linear feature extraction algorithms in the pursuit to identify the most adequate algorithm for the spike sorting problem.

## Non-linear feature extraction

The classical techniques for dimensinality reduction, such as PCA and MDS, are computationally eficient and perfectly able to find the structure of linear spaces (11). However, they encounter difficulties when non-linear structures are present (12). Non-linear manifold learning algorithms seek to discover a low-dimensional embedding (or a manifold) within the high-dimensional input data. These methods are capable of preserving the intristic geometry (including local neighbourshood and data topology) through their approximation of the underlying manifold, rather than relying on global linear projections such as PCA (13) (14).

In spike sorting, each detected spike waveform can be viewed as a high-dimensional vector. We can consider that the shapes of spikes vary from their ‘true shape’ due to recording artifacts. Therefore, non-linear manifold feature extraction techniques may be able to disentangle these factors by yielding embeddings that are robust to perturbations (15) and offer separability in overlapping clusters (generated by linear techniques). Moreover, modern manifold techniques have been designed to handle large volumes of data by employing sparse neighbourhoods graps and optimisation for scalability (16,17). This makes them a viable candidate for the spike sorting of high-density probes.

## The challenges of spike sorting [TBRewritten]

Spike sorting is inherently complex for several reasons. Brain recordings are inherently subject to the distortion of the spike waveforms due to the reasons specified above, these phenomena affecting spike shape generate clusters that do not have a well-defined separation boundary. This overlap of clusters is a struggle for most clustering algorithms, especially if coupled with data imbalance. Cluster imbalance in neuronal data appears from the variability in firing rate of neurons. Neuronal activity is dynamically modulated by neural circuits, causing individual cells to fire at widely varying rates (18) (19); this variability generates clusters of disparate sizes and yields an intrinsic imbalance in the dataset. Moreover, neuronal activity takes place on a millisecond timescale, thus even relatively short brain recording sessions can produce a vast quantity of data (3). In this context, single-unit activity refers to the spikes of one neuron that can be isolated as a single cluster, whereas spikes from more distant neurons typically appear with lower amplitudes (poor signal-to-noise ratio), and cannot be reliably separated (resulting in a single cluster being identified)—these are generally referred to as multiunit activity (5).

These non-linear manifold feature extraction methods often outperform linear feature spaces (20) and may be able to simultaneously denoise waveforms which can create dense clusters and increase the variability between the spikes of different neurons which can create separable clusters. In this study, we therefore evaluate a suite of representative non-linear feature extractors (e.g., Isomap, LLE, Spectral Embedding, Diffusion Maps, UMAP, TriMap) in comparison with traditional feature extraction methods and other non-linear feature extracton methods, to systematically compare how each manifold embedding influences cluster separability and spike-sorting performance across datasets.

The paper is structured as follows. Section 2 reviews traditional feature‐extraction techniques and their performance in spike sorting, outlines the methods proposed for spike sorting, and describes the datasets and evaluation metrics. In Section 3, we assess the methods across multiple metrics and offer a critical interpretation of their performance. Finally, Section 4 examines the limitations of the methods proposed for spike sorting and presents our concluding findings.

# Materials and Methods

## Feature extraction algorithms [TBRewritten]

The most important step of the spike sorting pipeline is the feature extraction where the high-dimensional space of the spike waveform is projected to a usually lower-dimensional space which contains the most informative features. The purpose is thus dual, to preserve as much as possible from the data structure of the original feature space in the reduced space and at the same time to reduce the space as much as possible. There are many criteria by which feature extraction methods may be categorized, such as convexity or linearity (21). Here, we separate the methods used into 3 categories: linear, non-linear, and non-linear manifold feature extraction methods.

### Linear feature extraction methods

The most common algorithm for feature extraction is the Principal Component Analysis (PCA) (13) and it has been throughly used in spike sorting (14) (5) as well. Even recently developed spike sorting pipelines employ PCA in their computations (31). PCA identifies orthogonal directions, or eigenvectors, based on maximum variance. PCA projects the original feature space into a new feature space, called principal components, based on the eigenvectors obtained through the eigendecomposition. Essentially, PCA rotates the coordinate system to align with maximum variance. Dimensionality reduction can be achieved by discarding components while preserving data variance, most commonly only the first two or three principal components represent 70-80% of the variance of the original feature space and only these are kept (29) (30). However, variance may not be the best approach for the separability of clusters (2) (5) as the discarded low-variance features may encode more information for separability.

Multidimensional Scaling (MDS) (42) creates low-dimensional representations that attempt to preserve the relationship between data points. Its classical version computes a distance matrix between all points to find coordinates in a lower-dimensional space that best match the original distances by minimizing a stress function. By preserving original distances, it can be considered a linear approach. For the Euclideean distance, MDS produces similar results to PCA.

Independent Component Analysis (ICA) (32) was designed to separate multivariate signals into independent components. Nevertheless, it was shown to be highly performant in the spike sorting domain as well (33) (34). In constrast to PCA which finds uncorrelated components, ICA seeks statistically independent sources by iteratively maximizing non-Gaussianity (using measures like kurtosis or negentropy). Thus, ICA works under the assumption that the signals are linear mixtures of non-Gaussian independent signals. The ICA algorithm effectively unmixes the signals by finding an unmixing matrix that produces the most statistically independent outputs.

### Non-linear feature extraction methods

Kernel PCA (KPCA) (36) is a non-linear extension of PCA through the use of the “kernel trick”. A non-linear kernel is utilised to map the input data into a possibly-higher dimensional feature space followed by PCA. Through the computation and the extraction of the eigenvectors of the kernel matrix (representing the inner product space), KPCA can capture non-linear relationship that PCA misses without the additional computation of coordinates in the higher-dimensional space.

A non-metric version of Multidimensional Scaling (MDS) (42) can preserve the ordering of distances rather than the values themselves. In other words, points that are closer than other in the original space are closer as well in the embeddings obtained. This is ahieved through the transformation of original space using a monotonic function and iteratively miniziming the same stress function. Through the ordering of distances, the non-metric MDS may be able to preserve the structure of data points when the relationship between similarity and distance is non-linear.

Self-Organizing Map (SOM) (44) creates a mapping between the data points and a two-dimensional grid where similar high-dimensional inputs are located nearby to each other. The grid of “neurons” is initialized in the low-dimensional space, followed by a training process where the input data is presented repeatedly to update the neuron (and the neighbourhood) that best matches the input. The competitive learning process of SOMs preserves the topology of the input data.

### Non-linear manifold feature extraction methods

Locally Linear Embedding (LLE) (38) preserves the local structure of data points by representing each as a weighted combination of its neighbors. It operates on the assumption that each neighborhood of points lies close on a locally linear patch of the manifold. The three steps of LLE are: identifying the k-nearest neighbours of each points, computing the weights which best reconstruct each point based on its neighbours (by solving linear system of equations) and finding a low-dimensional representation that preserve the reconstruction weights (by solving an eigenvalue problem).

Modified Locally Linear Embedding (MLLE) (39) is an extension of LLE which employs multiple weight vectors for each data point that obtain valid reconstructions. By employing alignment techniques, it allows for the identification of a global embedding that respects the constraints of each set of weight vectors.

Hessian-based Locally Linear Embedding (HLLE) (40) is another extension of LLE which use the Hessian operator to capture the local structure of the data. The Hessian matrix (representing the second derivative of the manifold) is computed for each neighborhood of nearest neighbours. HLLE identifies directions along which the manifold is locally flat by finding the null space of the Hessian. These directions form the basis for the low-dimensional embedding.

Local Tangent Space Alignment (LTSA) (41) is another extension of LLE which aligns local tangent spaces to caputre the global structure of the data. The tangent space is computed (as a linear approximation using principal components) for neighbourhoods as the k-nearest neighbours. The embedding is found by aligning these tangent spaces (by solving an eigenvalue problem).

KepplerMapper (46) creates a graph representation of the data by leveraging concepts from topological data analysis to preverse the shape and connectivity of the original data. It projects the data onto a filter function, such as a principal component. The projection if then divided into overlapping interval. These intervals are clustered and it creates a graph where nodes represent clusters and edeges represent overlap between clusters.

Isometric Mapping, or Isomap (12), attempts to maintain the geodesic distance between the data. Essentially, it flattens the manifold structure while preserving the geodesic distance. It constructs a graph connecting nodes to nearest neighbours. A distance matrix is computed by computing the shortest paths between pairs of points/nodes. It concludes by applying MDS to obtain the low-dimensional space.

T-distributed Stochastic Neighbor Embedding (t-SNE) (37) manages to create a lower-dimensional space by mapping high-dimensional data to lower dimensions through pairwise probability similarities while preserving both local and global structure. Gaussian distributions are used to compute the conditional probabilities that represent the similarities between the points in the original space. The t-distribution is used to computed the similarity probability distribution of the low-dimensional space. T-SNE minimizes the Kullback-Leibler divergence between input feature space and the reduced feature space by using the two distributions.

Spectral embedding (15) constructs a weighted graph representing the data and uses its Laplacian matrix for dimensionality reduction to preserve local structures. The graph connects each points to its nearby points, computes the graph Laplacian matrix, and finds its eigenvectors corresponding to the smallest non-zero eigenvalues. These eigenvectors form the low-dimensional embedding.

PHATE (45) models diffusion processes through heat kernels to create an embedding that captures the intrinstic structure of the data. A neighborhood graph is constructed and local affinities are computed. Diffusion is applied to capture multi-scale relationships, potential distances that preserve both local and global structure are computed and embedded into low dimensions using MDS.

Diffusion Map (43) uses diffusion processes on the manifold to capture the intrinsic structure of the data. A graph is constructed where edges represent the probability of transitioning between points in a random walk. The eigenvectors of the normalized graph Laplacian are computed, which correspond to different time scales of the diffusion process. These eigenvectors are the low-dimensional embedding which were created by preserving diffusion distances.

UMAP (17) maps the high-dimensional data into low-dimensional embeddings that have a similar topological structure. A weighted graph is constructed with edges representing k-nearest neighbours. The edge weights are assigned using a fuzzy set membership function based on distance, thus creating a fuzzy topological representation. A low-dimensional space is initialized (usually using spectral techniques) and a similar graph is created (with different edge weights). Finally, UMAP optimizes the low-dimensional space to minimize the cross-entropy between the two graphs.

TriMap (16) creates embeddings based on triplet constraints. These triplet constraints compare relative proximities between points. TriMap samples triplets of points *(i, j, k)* where *i* should be closer to *j* than to *k* in the embedding space. It then optimizes an objective function that minimizes violations of these constraints using gradient descent. By focusing on these relative proximity relations rather than absolute distances, TriMap efficiently captures both local and global structure.

## Clustering algorithms

External evaluation metrics require both the predicted cluster labels and the corresponding ground truth labels. As such, we employed K-Means (49) clustering to obtain the labels necessary for external metrics, applying it immediately after feature extraction. K-Means has long been used in clustering, with many adaptations developed over time. It was first utilized for spike sorting in 1988 (50) (51) and has remained the standard method for many years. Even recent spike sorting pipelines (8) (52) either rely on or are inspired by K-Means, and in a recent comparative study of 25 clustering methods, it demonstrated its continued competitive performance by ranking third (50).

K-Means (49) is a centroid-based clustering technique that achieves clustering by dividing the data space into *k* groups and allocating each point to the closest centroid according to Euclidean distance. But it has a number of drawbacks. First of all, it necessitates pre-specifying the number of clusters, which can be difficult for the real-world data. Nevertheless, there are preprocessing methods for finding the *k* parameter. Second, the method is non-deterministic in its most basic version, which means that different outcomes may be obtained from repeated executions. This problem has been resolved by more recent improvements that have increased its consistency. Thirdly, overlapping clusters are hard for K-Means to handle. However, this disadvantage is benefic for our analysis: the more performant feature extraction methods will improve cluster separation which will be shown by a higher K-Means performance.

## Performance metrics

Six metrics were used to evaluate the outcomes, the first three are external measures while the latter three serve as internal measures (48). They are: Adjusted Rand Index (ARI), Adjusted Mutual Information (AMI), V-Measure (VM), Calinski-Harabasz Score (CHS), Davies-Bouldin Score (DBS), and Silhouette Score (SS). Since spike sorting ends in clustering, these clustering metrics are suitable for evaluating the quality of feature extraction. External metrics measure how accurately the clustering algorithm recovers known class labels; they reflect how distinctly features separate the data—if features are perfectly separated then typically high external scores are achieved by the clustering algorithms. Internal metrics measure cluster compactness, separation, and shape as independent of ground truth; therefore they are appropriate for judging feature extraction when synthetic datasets provide true labels. Internal metrics describe cluster structure; external metrics measure agreement with true labels.

Using multiple measures instead of one index helps us take different aspects of clustering performance into account. Table 1 gives the intuitive meaning for each metric and its scoring range. A technique that does well across these varied measures suggests a more sound and even approach to feature extraction, lessening the chance of bias in the assessment.

**Table 1** – A short description of each performance evaluation metric, specifying its type and range.

|  |  |  |  |
| --- | --- | --- | --- |
| Name | Type | Description | Range [worst, best] |
| ARI | External | Chance‐corrected score based on pairwise comparisons of objects, rewarding when pairs are either consistently grouped together or separated in both predicted and ground‐truth clusters | [-1, 1] |
| AMI | External | Chance‐corrected score based on entropy‐based mutual information between the predicted and actual labels, with an adjustment for the expected value under random assignments | [0, 1] |
| Purity | External | Fraction of correctly assigned points across all clusters by assigning each cluster to the majority true class within it. | [0, 1] |
| DBS | Internal | Average similarity ratio of each cluster with its most similar cluster, where similarity is defined as the sum of within‐cluster scatter relative to between‐cluster separation. | (Inf, 0] |
| CHS | Internal | Ratio of between‐cluster dispersion to within‐cluster dispersion, normalized by the number of clusters and total points. | [0, Inf) |
| SS | Internal | Average across all data points for the normalized difference between its mean intra‐cluster distance and lowest mean inter‐cluster distance | [-1, 1] |

[TBRewritten]

ARI (53) (54) (55) (3) is an adjustment of the Rand Index (RI) metric in order to handle chances. ARI is an external clustering metric; therefore, it requires a ground truth for the dataset. RI (56) (2) makes comparisons between pairs of points to determine if it is an agreement, when the two points are in the same cluster for both the predicted and the true labels, or a disagreement, when they belong to different clusters. The formulas used to calculate the metric are the following:

where *MaxRI* is the upper bound and *ExpectedRI* is theexpected placement of pairs in the same class using the permutation model and calculated based on the contingency table (53).

AMI (54) (57) (5) is an adjustment of the Mutual Information (MI) metric through the use of entropy, denoted as H. Moreover, AMI also contains the normalization (54) (58) (59) of Normalized Mutual Information. MI (4) is calculated between two clusters U and V, where N is the size of the dataset and |X| is the number of points in subset X.

V-Measure (60) (6) is the harmonic mean of Homogeneity and Completeness. A cluster is considered to be homogeneous (7) when all the points of that cluster are part of the same class. By switching the predicted and true labels, completeness is obtained. Completeness (8) is achieved when all the points of a class are part of the same cluster. We have chosen beta equal to 1 as given by the original formula (60).

where H(C|K) is the conditional entropy of the true cluster given the predicted cluster, H(C) is the entropy of the true cluster, while H(K|C) is the conditional entropy of the predicted cluster given the true cluster and H(K) is the entropy of the predicted cluster.

All the metrics presented until this point are external metrics and require a ground truth to compare with the predicted labels. Furthermore, all these metrics have bounded scores in the [0, 1] interval with higher values being more desirable.

The following three metrics are internal and therefore do not require a ground truth to be used. The internal metrics were used with the ground truth labels for the evaluation of the synthetic datasets. These metrics evaluate the intra-cluster and inter-cluster distances and the morphology of the clusters producing an adequate evaluation of the feature extraction capabilities.

DBS (61) (62) (63) (10) finds the mean similarity between clusters, where similarity, denoted as *R* (9), is defined by the distance between clusters and their sizes. The minimum value of this index is 0. The closer the result is to 0, the better separation exists between clusters. This may come as counterintuitive as it is the only metric where lower values represent a higher performance. The DBS metric is given by the following equations:

where *si* is the mean of all distances between the points of cluster *i* and its centroid, *di,j* is the distance between clusters *i* and *j* given by their centroids, and *max(Ri,j)* is the maximum similarity of clusters *i* and *j*.

CHS (48) (60) (11), also known as Variance Ratio Criterion, calculates the ratio between the intra-cluster and inter-cluster dispersion. Where *tr(X)* denotes the trace of between cluster *Bk* or within-cluster *Wk* dispersion matrix, *n* denotes the size of the dataset and *k* the number of clusters. The dispersion is defined as the sum of squared distances. For this metric, a higher value indicates a better result.

SS (60) (64) (12) is calculated by measuring the mean distance between a point and the rest of the points of that cluster and the mean distance between the point and all the points of the nearest cluster. The score is bound between [-1, 1] where -1 represents an incorrect clustering, 0 overlapping clusters, and 1 a dense clustering. SS aims for the standard concept of a cluster, dense and well separated, therefore such cases will give a higher score. The equation of SS is the following:

where *b* denotes the average of all distances between a point in cluster *i* and all points of the nearest cluster *j*, and *a* the average of all distances between a point in cluster *i* and all other points in the same cluster.

It is important to mention that although used in evaluation of spike sorting techniques (65) (66), accuracy is not a suitable performance metric. First, because spike sorting is unsupervised and accuracy requires labels. Second, neuronal data is imbalanced because of the various firing rates of individual neurons, and it is has been extensively shown that accuracy is not appropriate for evaluating tasks on imbalanced data (67) (68) (69) (70). Nevertheless, through the use of the chosen metrics, we are able to evaluate the separation and shapes of cluster using the internal metrics and the correctness of clustering using the created features using the external metrics.

## Data [TBRewritten]

### Synthetic datasets

The validation of deep clustering methods was made by comparing the different methods with [\*\*\*PCA, ICA and Isomap\*\*\*]. The chosen datasets, 95 in number and denominated as simulations, originate from the Department of Engineering, University of Leicester UK and are publicly available. Each simulation is a dataset. The creation of these simulations was based on recordings from the neocortex of a monkey. They were generated using 594 different spike shapes (9). The original study that introduces the simulations (9) also reviews different clustering algorithms and their results. Out of 20 different units, these algorithms were able to detect 10 in the best case.

The datasets were generated based on a real dataset recorded “in vivo”. The waveform contains 316 points originally sampled at 96 KHz; afterwards this frequency was reduced to 24KHz, therefore 79 samples describe a spike. Being synthetic datasets, each of these spikes has a label, which allows for the use of external metrics to evaluate performance. Each simulation contains a multi-unit cluster, which is the noise, and a number of clusters that varies between 2 and 20. Each unique number of clusters has 5 simulations. Thus, there are 5 simulations with 2 clusters, 5 simulations with 3 clusters, and so on.

All but one of the clusters are single-units between 0 and 50μm away from the electrode. The firing rate follows a Poisson distribution with a mean between 0.1 and 2Hz. The amplitudes follow a normal distribution and have been scaled to values between 0.9 and 2 to simulate real data. No spikes with temporal overlapping are present in the data, such that spikes have at least 0.3ms between them.

The generated multi-unit cluster was added in order to increase the complexity of clustering for the tested algorithms. The simulated multi-unit contains 20 spike shapes, each of the 20 neurons firing being between 50-140μm away from the electrode. The amplitude of the spikes was fixed to 0.5, with an overall composite firing rate of 5Hz, with each of the 20 individual composing neurons having a firing rate mean of 0.25Hz following an independent Poisson distribution. Here, in order to increase clarity, the multi-unit cluster is always color-coded in white in all figures.

To evaluate the proposed approach in comparison with other state-of-the-art methods we have chosen the following 4 simulations out of the 95 available as they are representative of the issues that are present in feature extraction methods and allow for the evaluation of the methods on varying numbers of clusters covering a wide range and enabling a comprehensive evaluation of performance:

* Simulation 1 (Sim1 - Fig 3a), containing 16 single-unit clusters and a multi-unit cluster (in total 17) with 12012 samples.
* Simulation 4 (Sim4 - Fig 3b), containing 4 single-unit clusters and a multi-unit cluster (in total 5) with 5127 samples.
* Simulation 16 (Sim16 - Fig 3c), containing 8 single-unit clusters and a multi-unit cluster (in total 9) with 7556 samples.
* Simulation 35 (Sim35 - Fig 3d), containing 12 single-unit clusters and a multi-unit cluster (in total 13) with 9481 samples.
* Simulation 14, containing 3 single-unit clusters and a multi-unit cluster (in total 4) with 4507 samples. This dataset was used for the visualization of the impact of alignment on feature extraction in section 3.1.

These simulations can also be viewed in Fig 3 through the use of PCA to reduce the dimensionality from 79 to 2. The overlapping clusters produced by PCA can be clearly seen in Fig 3, in none of the datasets is it able to perfectly separate all clusters.

Scatter chart

Description automatically generated

**Fig 3. PCA projection of the synthetic datasets.** PCA projection of 4 different simulations with distinct numbers of clusters; the colors represent the cluster assignment given in the ground truth.

### Real datasets [TBRewritten]

The electrophysiological *“in vivo”* data was recorded from the brain of anaesthetized adult mice of the C57/B16 strain with A32-tet probes (NeuroNexus Technologies, Inc) at 32 kSamples /s (Multi Channel Systems MCS GmbH) during a visual stimulation. The stimuli were presented monocularly on a Beetronics 12VG3 12-inch monitor with a resolution of 1440x900, at 60fps and consisted of full-field drifting gratings (0.11 cycles/deg; 1.75 cycles/s; variable contrast 25–100%; 8 directions in steps of 45°). The animals, on which the extracellular activity was recorded, were placed in the stereotaxic holder (Stoelting Co, Illinois, United States) and anaesthetized. Anesthesia was induced and maintained with isoflurane (ISO) in oxygen (5% for induction, 1-3% for maintenance). The heart rate, respiration rate, core body temperature, and pedal reflex were constantly monitored. A circular craniotomy (1x1 mm) was performed over the left visual cortex of the animal centred on 0-0.5 mm anterior to lambda, 2-2.5 mm lateral to midline. To obtain multiunit activity (MUA) containing signals, the extracellular data was digitally filtered using a band-pass filter with a range of 300Hz-7000Hz using a bidirectional Butterworth IIR filter of order 3. An amplitude threshold, most commonly chosen between 3 and 5 (2) standard deviations of the recorded signal, was used to detect spike, which were then fed into the feature extraction algorithms. Spikes were identified as threshold crossings and subsequently used as input for the feature extraction algorithm.

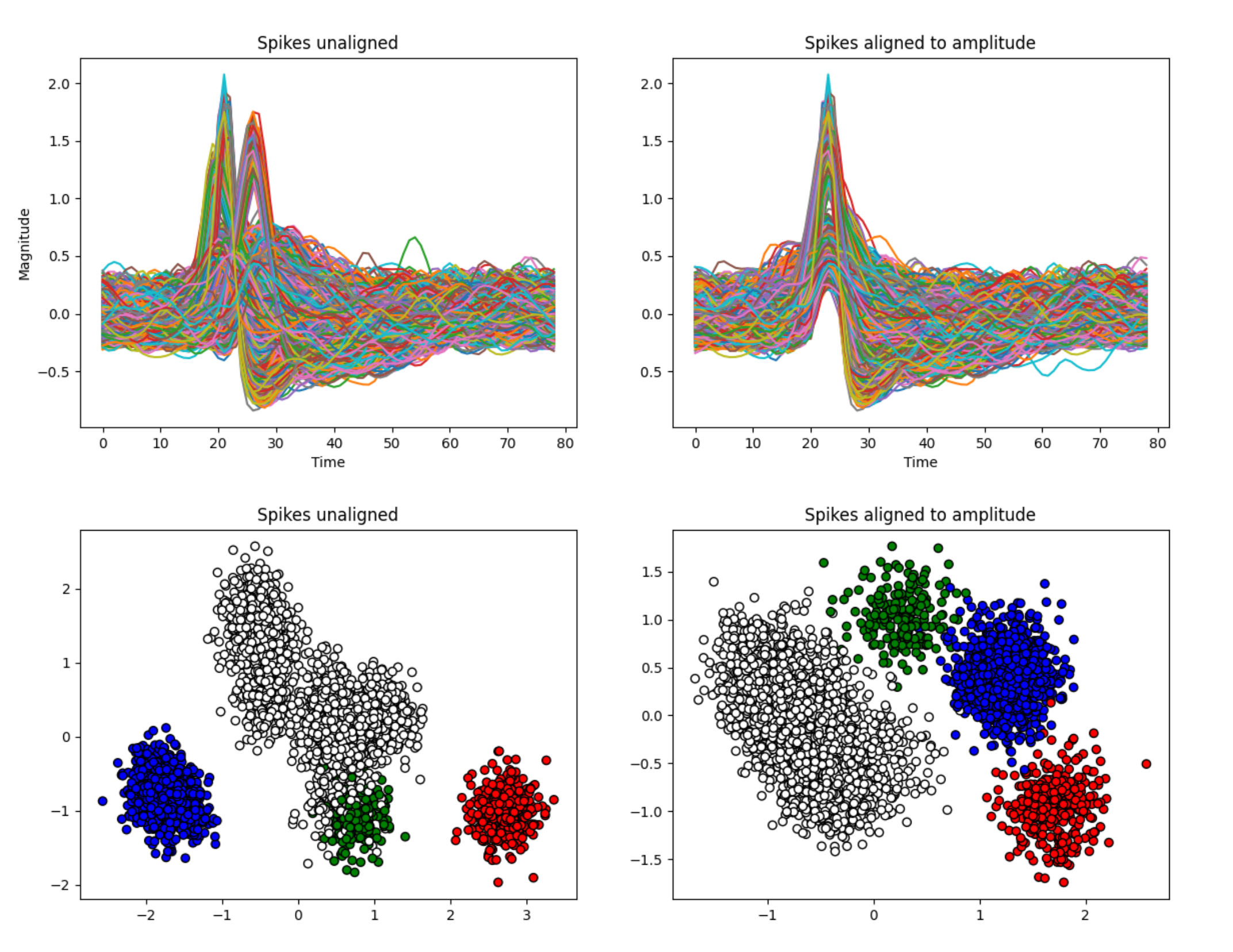
Multiple datasets were accumulated from each animal over a period of 4 to 6h in order to minimise animal use. All experiments were performed in accordance with the European Communities Council Directive of 22 September 2010 (2010/63/EU) and approved by the Local Ethics Committee (3/CE/02.11.2018) and the National Veterinary Authority (147/04.12.2018).

## Data preprocessing [TBRewritten]

Alignment has to be applied as a first step of preprocessing before the execution of the feature extraction method. We have used the following formula, for multiple types of alignment at a chosen index:

Naturally, the point of start of a portion of the samples has to be changed; this is indicated in formula (1) through the *new\_start* and *old\_start* terms. The *index* in equation (1) represents the point to which all spikes will be shifted. Thus, we can choose to align the maximum peak of all spikes to the average index of the maximum peak across all samples, as shown in Fig 4c. Another, better option, is to align the amplitudes to the middle of the sample as it provides information about the spike from the perspective of both pre- and post-amplitude. The *peak* in equation (1) represents the index at which the desired point of reference (typically, the peak) is found. For the alignment of the amplitude, it is the index of the maximum peak of each sample. The formula permits the alignment of any point of reference, such as the minimum peak (47).

In addition, we have applied two other preprocessing steps: scaling and shuffling.



**Fig 4. Impact of alignment.** PCA applied on Sim14 with and without alignment. The white cluster is kept together but the overlap with the blue cluster remains.

# Results [TBRewritten]

## Performance evaluation

### Performance evaluation of synthetic data [TBRewritten]

The 95 synthetic datasets (9) contain varying numbers of clusters and spike shapes providing the complexity required for a comprehensive evaluation of the methods. In Fig 5, we present the results obtained for each metric across all 95 datasets for each method presented. A statistical analysis using t-tests with a Bonferroni correction can be examined in the S4 Fig and a ranking of the methods based on their performance for each metric using Borda rank aggregation (71) in S2 Table.

### Performance evaluation of real data

# Discussion

# Bibliography

1. Carter M, Shieh J. Chapter 4 - Electrophysiology. In: Carter M, Shieh J, editors. Guide to Research Techniques in Neuroscience (Second Edition) [Internet]. San Diego: Academic Press; 2015 [cited 2022 Aug 2]. p. 89–115. Available from: https://www.sciencedirect.com/science/article/pii/B9780128005118000046

2. Quiroga RQ. Spike sorting. Scholarpedia. 2007 Dec 21;2(12):3583.

3. Bear M, Connors B, Paradiso M. Neuroscience: Exploring the brain: Fourth edition. 2015. 1 p.

4. Steinmetz NA, Aydin C, Lebedeva A, Okun M, Pachitariu M, Bauza M, et al. Neuropixels 2.0: A miniaturized high-density probe for stable, long-term brain recordings. Science. 2021 Apr 16;372(6539):eabf4588.

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

6. Stevenson IH, Kording KP. How advances in neural recording affect data analysis. Nat Neurosci. 2011 Feb;14(2):139–42.

7. Jun JJ, Steinmetz NA, Siegle JH, Denman DJ, Bauza M, Barbarits B, et al. Fully integrated silicon probes for high-density recording of neural activity. Nature. 2017 Nov;551(7679):232–6.

8. Pachitariu M, Steinmetz N, Kadir S, Carandini M, D HK. Kilosort: realtime spike-sorting for extracellular electrophysiology with hundreds of channels [Internet]. bioRxiv; 2016 [cited 2022 Jul 18]. p. 061481. Available from: https://www.biorxiv.org/content/10.1101/061481v1

9. 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.

10. Estivill-Castro V. Why so many clustering algorithms: a position paper. SIGKDD Explor Newsl. 2002 Jun 1;4(1):65–75.

11. Multivariate Analysis [Internet]. 1979 [cited 2025 May 4]. Available from: https://shop.elsevier.com/books/multivariate-analysis/mardia/978-0-08-057047-1

12. Tenenbaum JB, de Silva V, Langford JC. A global geometric framework for nonlinear dimensionality reduction. Science. 2000 Dec 22;290(5500):2319–23.

13. Mishra S, Sarkar U, Taraphder S, Datta S, Swain D, Saikhom R, et al. Principal Component Analysis. Int J Livest Res. 2017 Jan 1;1.

14. 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.

15. Belkin M, Niyogi P. Laplacian Eigenmaps for dimensionality reduction and data representation. Neural Comput. 2003 Jun 1;15(6):1373–96.

16. Amid E, Warmuth MK. TriMap: Large-scale Dimensionality Reduction Using Triplets [Internet]. arXiv; 2022 [cited 2025 May 2]. Available from: http://arxiv.org/abs/1910.00204

17. McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction [Internet]. arXiv; 2020 [cited 2025 May 2]. Available from: http://arxiv.org/abs/1802.03426

18. Buzsáki G. Rhythms of the Brain [Internet]. New York: Oxford University Press; 2006 [cited 2021 Dec 8]. 464 p. Available from: https://oxford.universitypressscholarship.com/10.1093/acprof:oso/9780195301069.001.0001/acprof-9780195301069

19. Lewicki MS. A review of methods for spike sorting: the detection and classification of neural action potentials. Netw Bristol Engl. 1998 Nov;9(4):R53-78.

20. Meilă M, Zhang H. Manifold learning: what, how, and why [Internet]. arXiv; 2023 [cited 2025 May 4]. Available from: http://arxiv.org/abs/2311.03757

21. Dimensionality reduction: a comparative review | BibSonomy [Internet]. [cited 2022 Aug 11]. Available from: https://www.bibsonomy.org/bibtex/2ed03568f0e9bca9cdaf6b25304e55940/peter.ralph

22. Meister M, Pine J, Baylor DA. Multi-neuronal signals from the retina: acquisition and analysis. J Neurosci Methods. 1994 Jan 1;51(1):95–106.

23. Pouzat C, Mazor O, Laurent G. Using noise signature to optimize spike-sorting and to assess neuronal classification quality. J Neurosci Methods. 2002 Dec 31;122(1):43–57.

24. Litke AM, Bezayiff N, Chichilnisky EJ, Cunningham W, Dabrowski W, Grillo AA, et al. What does the eye tell the brain?: Development of a system for the large-scale recording of retinal output activity. IEEE Trans Nucl Sci. 2004 Aug;51(4):1434–40.

25. Hulata E, Segev R, Ben-Jacob E. A method for spike sorting and detection based on wavelet packets and Shannon’s mutual information. J Neurosci Methods. 2002 May 30;117(1):1–12.

26. Chung JE, Magland JF, Barnett AH, Tolosa VM, Tooker AC, Lee KY, et al. A fully automated approach to spike sorting. Neuron. 2017 Sep 13;95(6):1381-1394.e6.

27. Ebbesen CL, Reifenstein ET, Tang Q, Burgalossi A, Ray S, Schreiber S, et al. Cell Type-Specific Differences in Spike Timing and Spike Shape in the Rat Parasubiculum and Superficial Medial Entorhinal Cortex. Cell Rep. 2016 Jul 14;16(4):1005–15.

28. Yuan Y, Yang C, Si J. The M-Sorter: An automatic and robust spike detection and classification system. J Neurosci Methods. 2012 Sep 30;210(2):281–90.

29. Glaser EM, Marks WB. ON-LINE SEPARATION OF INTERLEAVED NEURONAL PULSE SEQUENCES. In: Enslein K, editor. Data Acquisition and Processing in Biology and Medicine [Internet]. Pergamon; 1968 [cited 2022 Aug 11]. p. 137–56. Available from: https://www.sciencedirect.com/science/article/pii/B9780080035437500124

30. Abeles M, Goldstein MH. Multispike train analysis. Proc IEEE. 1977 May;65(5):762–73.

31. Toosi R, Akhaee MA, Dehaqani MRA. An automatic spike sorting algorithm based on adaptive spike detection and a mixture of skew-t distributions. Sci Rep. 2021 Jul 6;11(1):13925.

32. Hyvärinen A. Independent component analysis: recent advances. Philos Transact A Math Phys Eng Sci. 2013 Feb 13;371(1984):20110534.

33. Tiganj Z, Mboup M. Neural spike sorting using iterative ICA and a deflation-based approach. J Neural Eng. 2012 Dec;9(6):066002.

34. Lopes MV, Aguiar E, Santana E, Santana E, Barros AK. ICA feature extraction for spike sorting of single-channel records. In: 2013 ISSNIP Biosignals and Biorobotics Conference: Biosignals and Robotics for Better and Safer Living (BRC). 2013. p. 1–5.

35. Tharwat A, Gaber T, Ibrahim A, Hassanien AE. Linear discriminant analysis: A detailed tutorial. Ai Commun. 2017 May 16;30:169-190,.

36. Schölkopf B, Smola A, Müller KR. Kernel principal component analysis. In: Gerstner W, Germond A, Hasler M, Nicoud JD, editors. Artificial Neural Networks — ICANN’97. Berlin, Heidelberg: Springer; 1997. p. 583–8.

37. Zhou H, Wang F, Tao P. t-Distributed Stochastic Neighbor Embedding Method with the Least Information Loss for Macromolecular Simulations. J Chem Theory Comput. 2018 Nov 13;14(11):5499–510.

38. Roweis ST, Saul LK. Nonlinear Dimensionality Reduction by Locally Linear Embedding. Science. 2000 Dec 22;290(5500):2323–6.

39. Zhang Z, Wang J. MLLE: Modified Locally Linear Embedding Using Multiple Weights. In: Advances in Neural Information Processing Systems [Internet]. MIT Press; 2006 [cited 2025 May 2]. Available from: https://proceedings.neurips.cc/paper/2006/hash/fb2606a5068901da92473666256e6e5b-Abstract.html

40. Donoho DL, Grimes C. Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. Proc Natl Acad Sci. 2003 May 13;100(10):5591–6.

41. Zhang Z, Zha H. Principal Manifolds and Nonlinear Dimension Reduction via Local Tangent Space Alignment [Internet]. arXiv; 2002 [cited 2025 May 2]. Available from: http://arxiv.org/abs/cs/0212008

42. Borg I, Groenen PJF, editors. Constructing MDS Representations. In: Modern Multidimensional Scaling: Theory and Applications [Internet]. New York, NY: Springer; 2005 [cited 2025 May 2]. p. 19–35. Available from: https://doi.org/10.1007/0-387-28981-X\_2

43. Variable bandwidth diffusion kernels. Appl Comput Harmon Anal. 2016 Jan 1;40(1):68–96.

44. Kohonen T. Self-organized formation of topologically correct feature maps. Biol Cybern. 1982 Jan 1;43(1):59–69.

45. Moon KR, van Dijk D, Wang Z, Gigante S, Burkhardt DB, Chen WS, et al. Visualizing structure and transitions in high-dimensional biological data. Nat Biotechnol. 2019 Dec;37(12):1482–92.

46. Singh G, Memoli F, Carlsson G. Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition [Internet]. The Eurographics Association; 2007 [cited 2025 May 2]. Available from: https://doi.org/10.2312/SPBG/SPBG07/091-100

47. Dipalo M, Amin H, Lovato L, Moia F, Caprettini V, Messina G, et al. Intracellular and Extracellular Recording of Spontaneous Action Potentials in Mammalian Neurons and Cardiac Cells with 3D Plasmonic Nanoelectrodes. Nano Lett. 2017 May 23;17.

48. Rendón E, Abundez I, Arizmendi A, Quiroz EM. Internal versus External cluster validation indexes. 2011;5(1):8.

49. MacQueen J. Some methods for classification and analysis of multivariate observations. Proc Fifth Berkeley Symp Math Stat Probab Vol 1 Stat. 1967 Jan 1;5.1:281–98.

50. Veerabhadrappa R, Ul Hassan M, Zhang J, Bhatti A. Compatibility Evaluation of Clustering Algorithms for Contemporary Extracellular Neural Spike Sorting. Front Syst Neurosci [Internet]. 2020 [cited 2022 Jul 18];14. Available from: https://www.frontiersin.org/articles/10.3389/fnsys.2020.00034

51. Salganicoff M, Sarna M, Sax L, Gerstein GL. Unsupervised waveform classification for multi-neuron recordings: a real-time, software-based system. I. Algorithms and implementation. J Neurosci Methods. 1988 Oct;25(3):181–7.

52. 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.

53. Hubert L, Arabie P. Comparing partitions. J Classif. 1985 Dec 1;2(1):193–218.

54. Vinh NX, Epps J, Bailey J. Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance. :18.

55. Steinley D. Properties of the Hubert-Arable Adjusted Rand Index. Psychol Methods. 2004;9(3):386–96.

56. Fowlkes EB, Mallows CL. A Method for Comparing Two Hierarchical Clusterings. J Am Stat Assoc. 1983;78(383):553–69.

57. Strehl A, Ghosh J. Cluster Ensembles --- A Knowledge Reuse Framework for Combining Multiple Partitions. J Mach Learn Res. 2002;3(Dec):583–617.

58. Lazarenko D, Bonald T. Pairwise Adjusted Mutual Information. 2021.

59. Vinh N, Epps J, Bailey J. Information theoretic measures for clusterings comparison: Is a correction for chance necessary? ICML. 2009. 135 p.

60. Rosenberg A, Hirschberg J. V-Measure: A Conditional Entropy-Based External Cluster Evaluation Measure. In 2007. p. 410–20.

61. Caliński T, JA H. A Dendrite Method for Cluster Analysis. Commun Stat - Theory Methods. 1974 Jan 1;3:1–27.

62. Davies DL, Bouldin DW. A Cluster Separation Measure. IEEE Trans Pattern Anal Mach Intell. 1979 Apr;PAMI-1(2):224–7.

63. Halkidi M, Batistakis Y, Vazirgiannis M. On Clustering Validation Techniques. J Intell Inf Syst. 2001 Dec 1;17(2):107–45.

64. Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math. 1987 Nov 1;20:53–65.

65. Eom J, Park IY, Kim S, Jang H, Park S, Huh Y, et al. Deep-learned spike representations and sorting via an ensemble of auto-encoders. Neural Netw. 2021 Feb 1;134:131–42.

66. Radmanesh M, Rezaei AA, Jalili M, Hashemi A, Goudarzi MM. Online spike sorting via deep contractive autoencoder. Neural Netw [Internet]. 2022 Aug 5 [cited 2022 Aug 11]; Available from: https://www.sciencedirect.com/science/article/pii/S089360802200301X

67. Wegier W, Ksieniewicz P. Application of Imbalanced Data Classification Quality Metrics as Weighting Methods of the Ensemble Data Stream Classification Algorithms. Entropy Basel Switz. 2020 Jul 31;22(8):E849.

68. Sun Y, Wong AKC, Kamel MS. Classification of imbalanced data: a review. Int J Pattern Recognit Artif Intell. 2009 Jun;23(04):687–719.

69. Joshi MV, Kumar V, Agarwal RC. Evaluating boosting algorithms to classify rare classes: comparison and improvements. In: Proceedings 2001 IEEE International Conference on Data Mining. 2001. p. 257–64.

70. Weiss GM. Mining with rarity: a unifying framework. ACM SIGKDD Explor Newsl. 2004 Jun;6(1):7–19.

71. Dwork C, Kumar R, Naor M, Sivakumar D. Rank aggregation methods for the Web. In: Proceedings of the 10th international conference on World Wide Web [Internet]. New York, NY, USA: Association for Computing Machinery; 2001 [cited 2022 Dec 6]. p. 613–22. (WWW ’01). Available from: https://doi.org/10.1145/371920.372165