Article

A Comprehensive Exploration of Unsupervised Classification in Spike Sorting: A Case Study on Macaque Monkey and Human Pancreatic Signals

Francisco Javier Iñiguez-Lomeli [1], Edgar Eliseo Franco-Ortiz [1], Ana Maria Silvia Gonzalez-Acosta [2], Andres Amador Garcia-Granada [3] and Horacio Rostro-Gonzalez [2,3,*]

[1] Department of Electronics Engineering, University of Guanajuato, Carretera Salamanca—Valle de Santiago km 3.5 + 1.8 km, Salamanca 36885, Mexico; fj.iniguezlomeli@ugto.mx (F.J.I.-L.); edgar.franco@ugto.mx (E.E.F.-O.)

[2] Laboratory of Robotics, Bio-Inspired Systems and Artificial Intelligence, Faculty of Biological Systems and Technological Innovations, Benito Juárez Autonomous University of Oaxaca, Av. Universidad S/N. Ex-Hacienda 5 Señores, Oaxaca 68120, Mexico; goaa800904.fmc@uabjo.mx

[3] School of Engineering, Instituto Químico de Sarrià, Universitat Ramon Llull, 08017 Barcelona, Spain; andres.garcia@iqs.url.edu

* Correspondence: horacio.rostro@iqs.url.edu

Abstract: Spike sorting, an indispensable process in the analysis of neural biosignals, aims to segregate individual action potentials from mixed recordings. This study delves into a comprehensive investigation of diverse unsupervised classification algorithms, some of which, to the best of our knowledge, have not previously been used for spike sorting. The methods encompass Principal Component Analysis (PCA), K-means, Self-Organizing Maps (SOMs), and hierarchical clustering. The research draws insights from both macaque monkey and human pancreatic signals, providing a holistic evaluation across species. Our research has focused on the utilization of the aforementioned methods for the sorting of 327 detected spikes within an in vivo signal of a macaque monkey, as well as 386 detected spikes within an in vitro signal of a human pancreas. This classification process was carried out by extracting statistical features from these spikes. We initiated our analysis with K-means, employing both unmodified and normalized versions of the features. To enhance the performance of this algorithm, we also employed Principal Component Analysis (PCA) to reduce the dimensionality of the data, thereby leading to more distinct groupings as identified by the K-means algorithm. Furthermore, two additional techniques, namely hierarchical clustering and Self-Organizing Maps, have also undergone exploration and have demonstrated favorable outcomes for both signal types. Across all scenarios, a consistent observation emerged: the identification of six distinctive groups of spikes, each characterized by distinct shapes, within both signal sets. In this regard, we meticulously present and thoroughly analyze the experimental outcomes yielded by each of the employed algorithms. This comprehensive presentation and discussion encapsulate the nuances, patterns, and insights uncovered by these algorithms across our data. By delving into the specifics of these results, we aim to provide a nuanced understanding of the efficacy and performance of each algorithm in the context of spike sorting.

Keywords: biosignals; unsupervised classification; spike sorting; PCA; K-means; self-organizing maps; hierarchical clustering

1. Introduction

The realm of neural signal analysis has witnessed remarkable advancements with the emergence of clustering methods that play a crucial role in uncovering intricate patterns within multi-dimensional data [1–3]. Among these techniques, spike sorting is crucial, facilitating the isolation and classification of individual action potentials from complex recordings [4–6].

Accurate spike sorting is fundamental for unveiling neural dynamics, understanding network interactions, and deciphering the underlying mechanisms governing neural behavior [7]. However, the extraction of action potentials from biosignals has presented formidable challenges due to the pervasive presence of noise, interference, and the diversity of action potential shapes in biological recordings [8–10]. In recent years, these main drawbacks have been studied from a computational viewpoint, where two seminal works have been published [11,12]. In these articles, the most commonly used techniques for biosignal analysis and spike sorting are mentioned, highlighting algorithms such as Wavelets and PCA as feature extraction techniques. Some newer algorithms, such as Support Vector Machines and neural networks, are also mentioned [13,14]. In the particular case of PCA, three papers mention that this technique has been used to reduce the dimensionality of the data. For clustering, some methods such as Bayesian and hierarchical clustering and Gaussian mixtures are also mentioned as a possibility to analyze this kind of data [15]. The methodologies reviewed in these two papers leverage the synergistic potential of machine learning, in particular, clustering methods and signal processing, to unravel the complexities of action potential detection and classification.

Recent articles have introduced novel methods for spike sorting: In [16], they introduced an approach for spike sorting by using a one-dimensional convolutional neural network; their method achieved high accuracy in synthetic data. However, it has not been tested with real acquisitions that have varying noise levels and overlapping spikes. They outperformed methods like WMsorting and deep learning multi-layer perceptron (MLP) models. The study by [17] proposed a deep learning-based technique for spike sorting, termed deep spike detection (DSD), to improve spike detection accuracy. DSD incorporates two convolutional neural networks (CNNs) into the conventional spike sorting pipeline. These CNNs are employed for selecting active neural channels and removing artifacts from the chosen channels, enhancing the overall classification performance. In [18], they present another approach to spike sorting using a deep learning-based autoencoder method. This method aims to improve clustering performance compared to traditional techniques like PCA and ICA. The improvement has been demonstrated using both synthetic and real datasets. Autoencoders are capable of learning underlying features from unlabeled data. Spike sorting involves extracting informative features from spike waveforms, and autoencoders can discover these features in an unsupervised manner, eliminating the need for labeled data, which can be difficult to obtain in spike sorting. Additionally, autoencoders have been shown to be robust to noise, which is a common challenge in spike sorting due to the low signal-to-noise ratio of recordings. In [19], the authors proposed a deep learning network based on a convolutional neural network and sliding window Long Short-Term Memory for spike sorting. LSTM [20] is a recurrent neural network architecture that has recently attracted attention in the study of biosignals due to its ability to process spatiotemporal information. This paper presents important results for low noise and simulated data, which is not the case for real neural recordings. In the same way as reinforcement learning, Liu et al. [21] proposed a method for sorting under multi-class imbalance in two public datasets with real neural recordings. They proposed a Markov sequence decision and constructed a dynamic reward function (DRF) to improve the sensitivity of the agent to minor classes based on the inter-class imbalance ratios. Another paper using deep learning was presented in [22], where the authors developed a deep learning method, which learns contextualized, temporal, and spatial patterns and classifies them as channels containing neural spike data or only noise. From this, they created a batch of waveforms to detect spikes in data recorded from a single tetraplegic patient. From the same author, in [14], a method based on K-means for classifying the detected spikes was presented. The methodology employed by these authors aligns closely with our approach, as we share a parallel trajectory. Much like their sequence, we initially published a paper focusing on spike detection and shifted our focus toward classifying the detected spikes. However, in our instance, our exploration delves into distinct clustering algorithms. The K-means algorithm was also recently investigated with template optimization in phase space as

demonstrated in [8]. This method obtained interesting results on both simulated data and real neural recordings. Wavelet [23] is another approach widely used in the analysis of biosignals. In this paper, the authors use a continuous wavelet transform (CWT) with optimized parameters to sort artificial and real data. The authors suggest that their results outperform those based on PCA.

Several notable and contemporary contributions within the realm of spike sorting have been highlighted [24]. Nevertheless, providing an exhaustive review of all existing literature in this domain is an intricate task, further compounded by the challenges of conducting a direct comparison across this large number of works. The intricacy arises due to the divergent employment of various databases and algorithms for detecting and classifying potentials.

In this context, we have spotlighted articles resembling our proposition, particularly those grounded in PCA and K-means methodologies. However, to the best of our knowledge, our pursuit did not yield evidence of articles employing hierarchical clustering or Self-Organizing Maps as classification methods in neural recordings. Even those employing PCA and K-means do so in a manner distinct from the approach proposed in this study. Specifically, our approach involves an initial step of utilizing spike detections derived from a prior research endeavor conducted by the research group and documented in [25]. This antecedent work employed an adaptive threshold method for real-time action potential detection. This precursor dataset facilitated the derivation of feature matrices capturing statistical attributes of the action potentials. These matrices were then subjected to three distinct clustering algorithms: K-means, hierarchical clustering, and Self-Organizing Maps (SOMs). Notably, within the specific context of K-means, PCA was integrated to reduce dimensionality and enhance the overall outcomes.

The article is structured as follows: First, the introduction is presented in Section 1. Subsequently, the research methodology developed in this article is outlined in Section 2. The biosignals and spike detection are covered in Section 2.1. Unsupervised classification methods and results are presented in Section 3. Finally, the discussion and the conclusions are drawn in Sections 4 and 5, respectively.

2. Materials and Methods

Spike sorting typically encompasses two distinct stages: detecting action potentials and their subsequent classification. Each of these stages is further divided into sub-stages. The first stage involves initial processing steps, such as filtering, while the second stage involves extracting pertinent features for classification. Although these two stages are intrinsically interconnected, innovation does not necessarily unfold in both simultaneously. In this context, the current study extends the prior research undertaken by the same research group. In a previous endeavor detailed in [25], our group developed a real-time action potential detection hardware implementation for the two distinct biosignals used in this work, a macaque monkey signal and a human pancreatic signal. In that study, the research team successfully devised an FPGA implementation of an adaptive threshold method tailored for detecting action potentials (spikes).

What is now proposed is to classify these spikes using various clustering methods, among which the use of Self-Organizing Maps stands out. To the best of our knowledge, SOM has not been applied to classifying these types of signals. We also propose using a combination of Principal Component Analysis and K-means, which have been independently used in other works. Finally, the implementation of hierarchical clustering is suggested, a technique which, according to our literature review, has also not been utilized in these type of signals.

The specific proposal is depicted at the bottom of Figure 1. We assume all detected spikes have been stored, from which we extract their statistical attributes. These attributes are subsequently normalized and serve as inputs for the unsupervised classification methods to execute their clustering procedures. Principal Component Analysis is also used to reduce the dimensionality of the data.
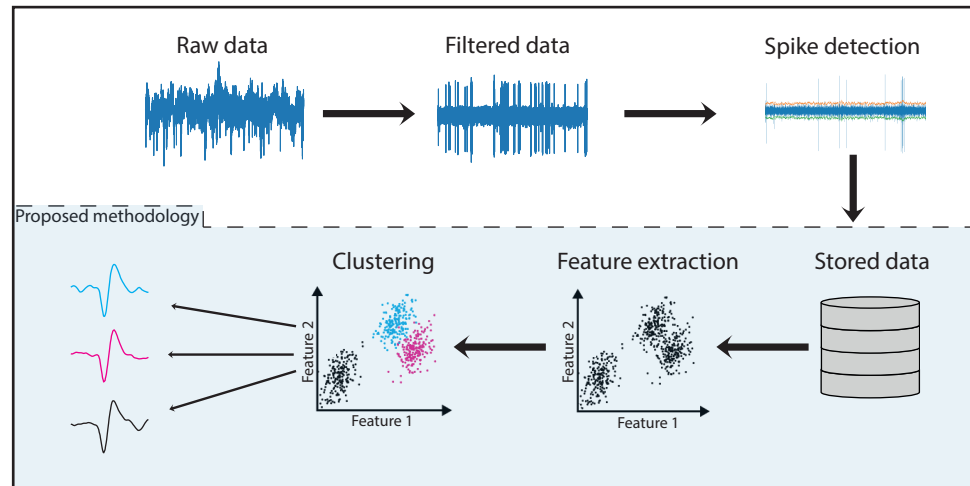
Figure 1. Methodology.

2.1. Biosignals and Spike Detection

In this study, two distinct biosignals are utilized: the first, captured in vivo from a macaque monkey (Figure 2a), has a duration of 559 s, from which only 25 s were used in this work, and a sampling frequency of 40 kHz; the second type, acquired in vitro for 13 s from human pancreatic cells (Figure 2b), is characterized by a sampling frequency of 10 kHz. The acquisition of these signals was not carried out in this work.
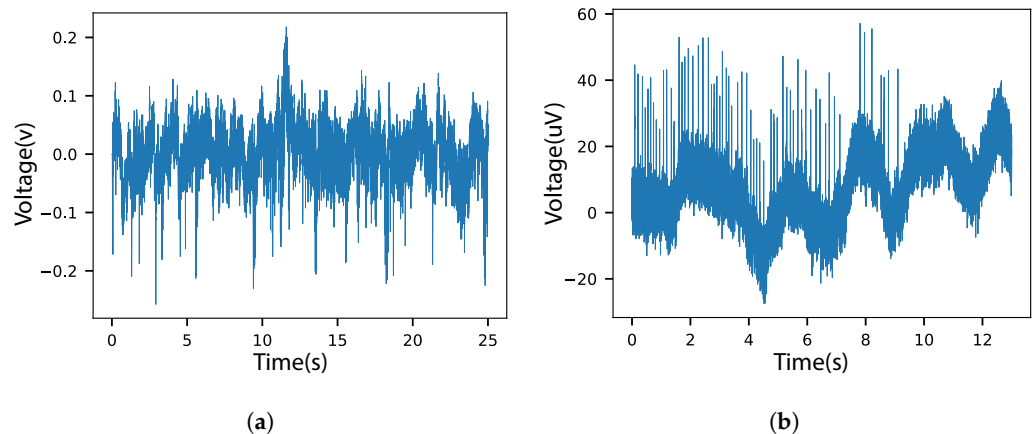

(**a**)

(**b**)

Figure 2. Samples of raw biosignals: (**a**) macaque monkey signal, (**b**) human pancreatic signal

Spikes are identified within the signals illustrated in Figure 2, showcasing specific instances of the raw biosignals acquired in vivo from a macaque monkey and in vitro from a human pancreas (see Figure A1 for some of the spike shapes found in the macaque monkey and human pancreatic biosignals). The spike detection process employs an adaptive thresholding method [26], automating the procedure while dynamically adjusting the threshold to remain above the signal's background noise level. Initial spike detection was performed in a previous study [25]. Our primary emphasis in this current research lies in thoroughly analyzing these spikes and classifying them.

Table 1 shows a subset of the statistical data extracted from 12 detected spikes within the monkey signal. A similar table exists for the spikes in the human pancreatic signal. In total, 327 spikes were identified within the monkey signals, while 386 spikes were detected in the human pancreatic signals. These data were subsequently utilized as inputs for clustering algorithms, enabling the aggregation of spikes exhibiting similar statistical characteristics. The table employs a color scale to indicate value intensity, revealing a notable level of variability within the dataset. This inherent variability has been addressed within this study through data normalization as follows:

$$z = \frac{x - \mu}{s} \tag{1}$$

Table 1. Feature matrix for a subset of 12 detected spikes within the monkey signal without normalization. The complete feature matrix contains all 327 detected spikes.

	Average Deviation	Variance	Wavelength	RMS
0	0.00849887	0.000187308	0.179557	0.0135544
1	0.0065741	7.09735×10^{-5}	0.0791692	0.0084606
2	0.00466847	3.33185×10^{-5}	0.0555136	0.0058420
3	0.00888798	0.000247485	0.1864338	0.0156036
4	0.00582276	3.57568×10^{-5}	0.0469245	0.0060722
5	0.0043362	2.65052×10^{-5}	0.0528824	0.0051284
6	0.00344435	2.10582×10^{-5}	0.0561784	0.0045793
7	0.00493412	2.81377×10^{-5}	0.0615169	0.0057737
8	0.00467569	3.12088×10^{-5}	0.0561566	0.0055327
9	0.00974478	0.00025559	0.1881129	0.0158705
10	0.00392492	2.48305×10^{-5}	0.0521927	0.0049249
11	0.00998542	0.00024017	0.1824477	0.0153818
12	0.00390623	2.15412×10^{-5}	0.0936657	0.0047380

This normalization primarily involves extracting the mean and scaling the data based on the standard deviation. This process aims to center the features and align the standard deviations, thereby mitigating the potential dominance of features with larger amplitudes within the clustering algorithms. This precaution prevents any feature from overpowering the learning process, ensuring that the estimator effectively learns from all features.

Upon normalizing the feature matrix presented in Table 1, the resulting data are illustrated in Table 2 for macaque monkey data and similarly for the human pancreatic data in Table 3. Unlike the former data, it is evident that the values are no longer widely separated, yet they maintain the same underlying distribution. Consequently, the clustering algorithms can effectively learn from all the features without the undue influence of any particular feature on the learning process.

Table 2. Feature matrix for a subset of 12 detected spikes in the monkey signal with normalization.

	Average Deviation	Variance	Wavelength	RMS
0	0.616557	0.507503	0.982358	0.690902
1	-0.0435587	-0.476921	-0.5054	-0.269358
2	-0.697112	-0.795558	-0.855978	-0.762997
3	0.750005	1.01672	1.08427	1.0772
4	-0.301236	-0.774925	-0.983269	-0.719604
5	-0.811067	-0.853212	-0.894972	-0.89753
6	-1.11693	-0.899305	-0.846125	-1.00105
7	-0.606004	-0.839398	-0.767008	-0.775876
8	-0.694634	-0.81341	-0.846448	-0.821307
9	1.04385	1.08531	1.10916	1.1275
10	-0.952118	-0.867384	-0.905193	-0.935889
11	1.12638	0.954852	1.0252	1.03539
12	-0.958528	-0.895218	-0.290559	-0.970757

The datasets presented in Tables 1–3 constitute the feature matrices used in the subsequent description of the clustering algorithms. These algorithms have undergone validation for both feature matrices, and the outcomes of these validations are presented in the subsequent sections.

Table 3. Feature matrix for a subset of 12 detected spikes in the human pancreatic signal with normalization.

	Average Deviation	Variance	Wavelength	RMS
0	−0.625955	−0.556876	−0.512561	−0.574781
1	−0.643811	−0.578161	−0.856704	−0.641413
2	1.92222	2.23712	1.76218	2.07959
3	1.72554	1.86247	1.81155	1.87429
4	−0.207561	−0.403196	0.215127	−0.250521
5	−0.689361	−0.589357	−0.879274	−0.674468
6	−0.468224	−0.508807	0.36071	−0.464447
7	1.92925	2.01965	1.38249	1.95593
8	1.92233	1.88622	1.43699	1.87117
9	2.01004	2.13762	1.93265	2.02572
10	−0.629296	−0.574017	−0.682783	−0.629894
11	−0.678137	−0.588473	−0.756864	−0.666986
12	0.800711	0.448739	0.658404	0.799653

2.2. Unsupervised Classification Methods

Three distinct methods were employed to classify the identified spikes for both signals: K-means, hierarchical clustering, and Self-Organizing Maps (SOMs). The protocol and parameters used during experiments are depicted in Figure 3. In this work, we chose to avoid using features incurring high computational costs. Therefore, we only utilized time-domain statistics as features. Below are the features utilized during the experiments:

- Mean Absolute Value (*MAV*). It is the average of the absolute values of the *N* samples taken:

$$MAV = \frac{1}{N} \sum_{i=1}^{N} |s(i)| \tag{2}$$

- Variance (*VAR*). It defines the dispersion of the data with respect to the mean:

$$VAR = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})^2 \tag{3}$$

- Root Mean Square (*RMS*) is a scalar value corresponding to the root mean square. It is used to help obtain the amplitude of the signal:

$$RMS = \sqrt{\frac{1}{N} \sum_{i=1}^{N} x_i^2} \tag{4}$$

- The Integral of the Absolute Value (*IAV*). It represents the sum of the absolute values of the signal in a period of time:

$$IAV = \sum_{i=1}^{N} |x_i| \tag{5}$$

- Simple Square Integral (*SSI*). It calculates the energy of the signal:

$$SSI = \sum_{i=0}^{N} |x_i^2| \tag{6}$$

Waveform Length (*WL*). It specifies the cumulative length of the waveform shape in a particular segment:

$$WL = \sum_{i=1}^{N-1} |x_{i+1} - x_i| \tag{7}$$

- Entropy (*H*). Entropy measures the complexity of an uncertain system. It is a statistical method for quantifying the unpredictability of variations in both deterministic and stochastic signals:

$$H = -\sum_{i=1}^{N} p_i ln(p_i) \qquad (8)$$

where p_i is the corresponding probability of each of the *N*-states.
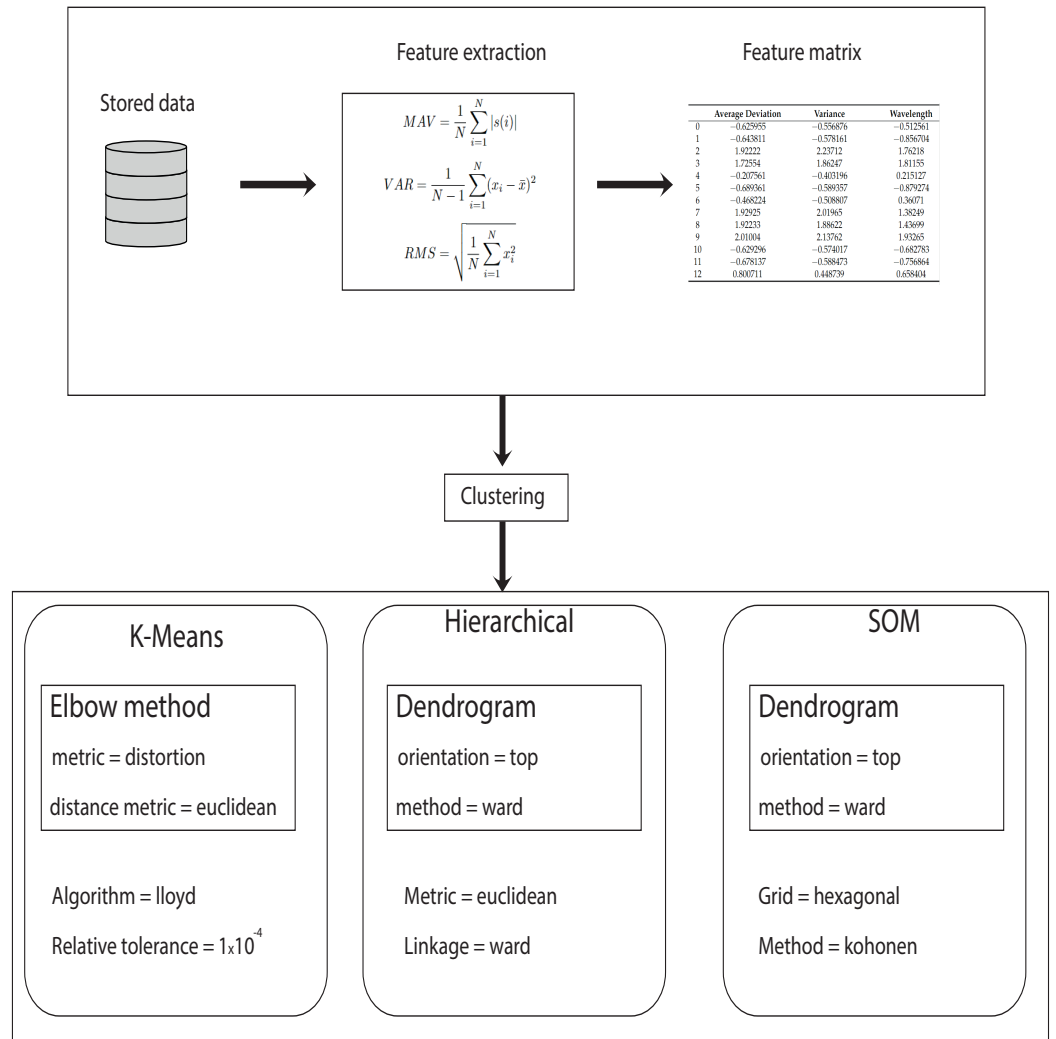
Figure 3. Protocol and parameters used during experiments with each unsupervised classification method.

To enhance the differentiation between action potentials, additional features based on their waveform characteristics were calculated. These features included the zero-crossing rate (the number of times the waveform crosses zero), the peak-to-peak distance (the distance between samples at the minimum and maximum peaks), and the peak amplitude. This focus on waveform properties aimed to improve the ability to distinguish between action potentials. The following sections delineate the outcomes attained through the application of these methodologies.

3. Experimental Results

3.1. K-means

The K-means algorithm [27,28] underwent testing through two approaches. Firstly, it was applied to the feature matrix acquired without data modifications or normalization as depicted in Table 1.

The elbow method was employed to ascertain the optimal number of clusters in the K-means algorithm. This involved calculating the dispersion curve of the data while varying the K value from 2 to 12. The resulting graphs of the elbow method for each signal before normalization can be observed in Figure 4a,b.

| (a) | (b) |
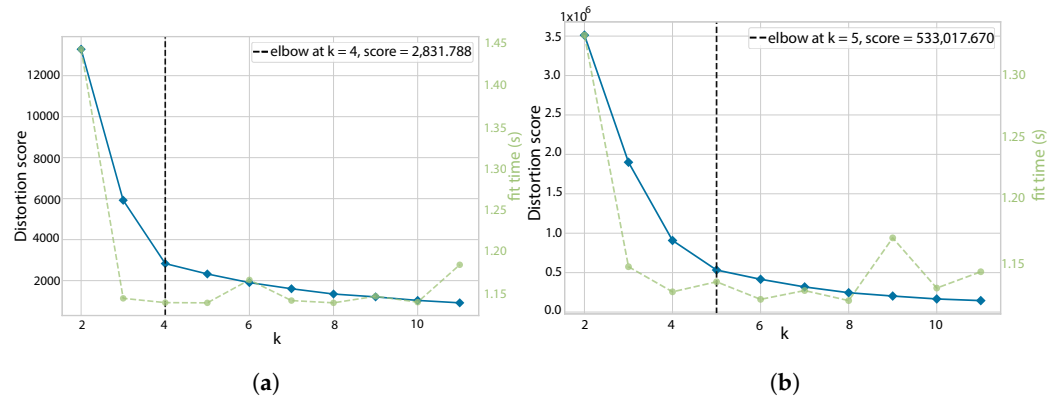
Figure 4. Visualization of the elbow method encompassing values of K ranging from 2 to 12, employed by the K-means algorithm for the clustering of detected spikes in (**a**) signals from monkeys and (**b**) signals from human pancreatic cells. The green line represents the time taken to train the clustering model for each cluster number, while the blue line represents the distortion score.

The elbow method graphs depicted in Figure 4 are generated by summing the squared distances between each point and their assigned center. They also indicate the training time for each value of K, represented by the green graph. Figure 4 illustrates that the optimal fit of K for monkey and human pancreatic detected spikes without normalization occurs at four and five, respectively.

Figure A2 shows the data distribution in blue before normalization with the centroids obtained with the K-means algorithm. The number of groups/clusters used was four for the spikes in the monkey signal and five for the spikes in the human pancreatic signal. It can be seen that for the clusters associated with the spikes in the monkey signal (Figure A2a), three centroids are very close at the bottom of the data, leaving only one centroid for the data at the top. Regarding the clusters associated with the spikes in the human pancreas signal (Figure A2b), two clusters overlap at the bottom of the graph. Although this grouping already gives us a previous idea of what is expected to be obtained, there is still a significant loss of information since all the data with greater amplitude are grouped in a single group.

To improve the outcomes achieved through K-means without normalization, a secondary approach is adopted. This approach involves utilizing the normalized data from Tables 2 and 3 as the input feature matrices for the K-means algorithm. Results for normalized data are shown in Figure 5.

Figure 5 displays the updated elbow graph and cluster distributions based on normalized data. Notably, the number of clusters for spikes within the monkey signal has been adjusted from four to six. Conversely, the number of clusters for spikes within the human pancreatic signals remains consistent, yet the distribution and differentiation among these clusters have been modified.

The application of K-means clustering yields Tables 4 and 5 for spikes in the monkey signal and Table 6 for spikes in the human pancreatic signal, revealing the spike shapes attributed to each group within the spikes distribution space. These tables provide insight into the number of spikes sorted within each group. Tables 4 and 5 correspond to the clustering before and after normalization for the spikes in the macaque monkey data, respectively. The latest shows three groups, which were generated from group 0 in Table 4 after data normalization.

(**a**)　　　　　　　　　　　　　　(**b**)

(**c**)　　　　　　　　　　　　　　(**d**)
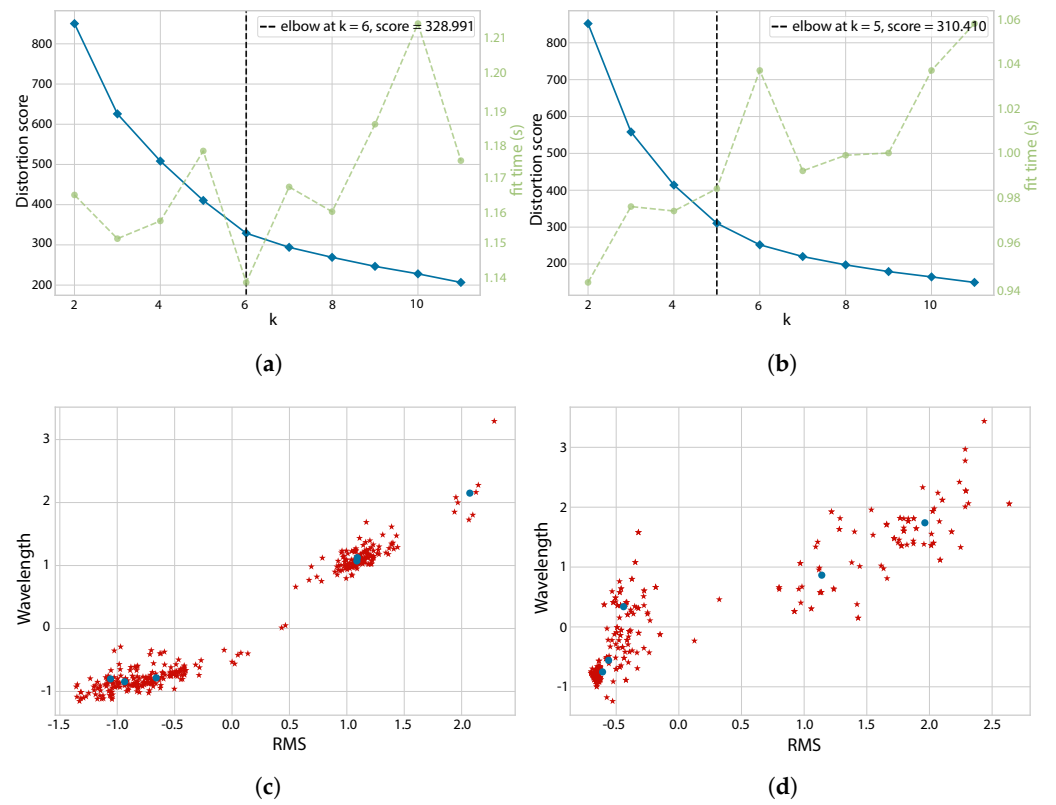
Figure 5. K-means clustering with normalized data: (**a**) elbow graph for spikes in the monkey signal, (**b**) elbow graph for spikes in the human pancreatic signal, (**c**) clusters/groups for the spikes in the monkey signal, and (**d**) clusters/groups for the spikes in the human pancreatic signal.

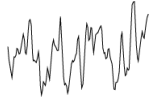**Table 4.** Spikes groups formed with K-means for the monkey signal (before normalization).

Group	Spike Shape	Number of Spikes
0		147
1		113
2		24
3		43

Table 5. Spikes groups formed with K-means for the monkey signal (after normalization).
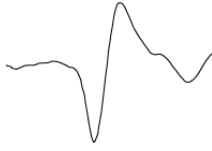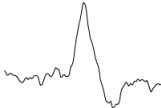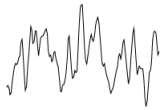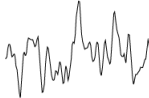
Group	Spike Shape	Number of Spikes
0-0		8
0-1		59
0-2		67

Table 6. Spikes groups formed with K-means for the human pancreatic signal (after normalization).

Group	Spike Shape	Number of Spikes
0		52
1		62
2		101
3		116
4		55

Although these action potentials have similar amplitude and shape, some differences can be perceived between them, mainly after the repolarization stage of the potential. Figure 6 shows the differences in these three spikes.
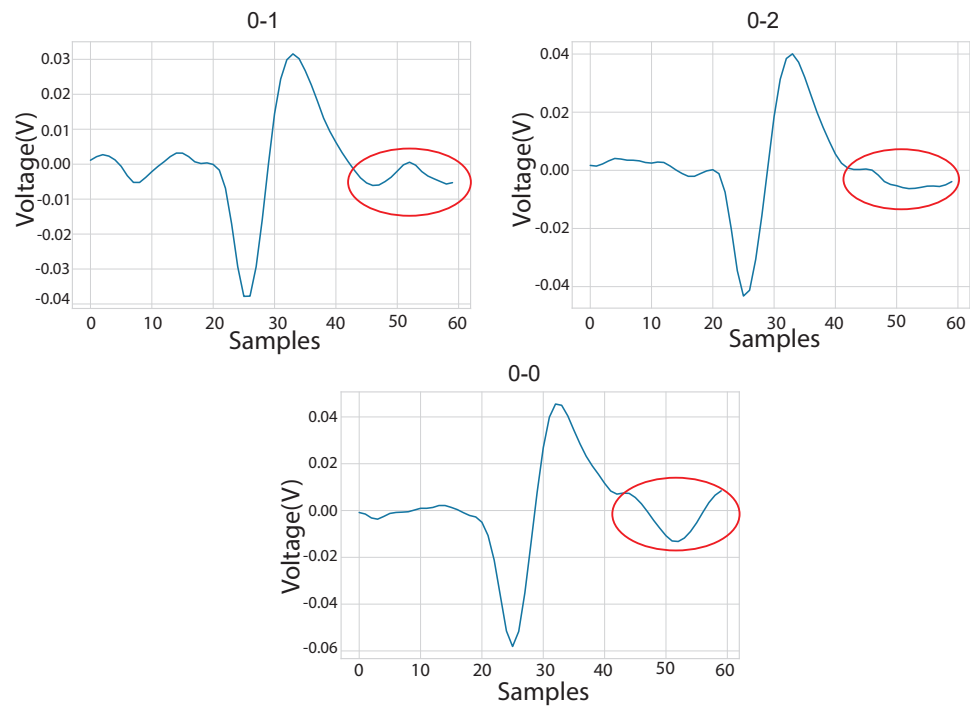
Figure 6. Differences of the spike shapes in Table 5.

Figure 7 presents a three-dimensional representation illustrating the distinctive spike shapes within each group formed from the detected spikes within the monkey signal. In this visualization, the three spike types outlined in Figure 6 are prominently discernible. Moreover, other detected shapes from the lower groups are also depicted, revealing a range of spike patterns. Notably, while certain shapes correspond to noise eliminated during the potential detection stage, others signify spikes resembling action potentials. However, the clustering method encounters challenges in distinguishing these from the noise. This confusion arises because some of these spikes share the same amplitude level as the detected noise, underscoring the complexity of isolating genuine action potentials in the presence of signal artifacts.
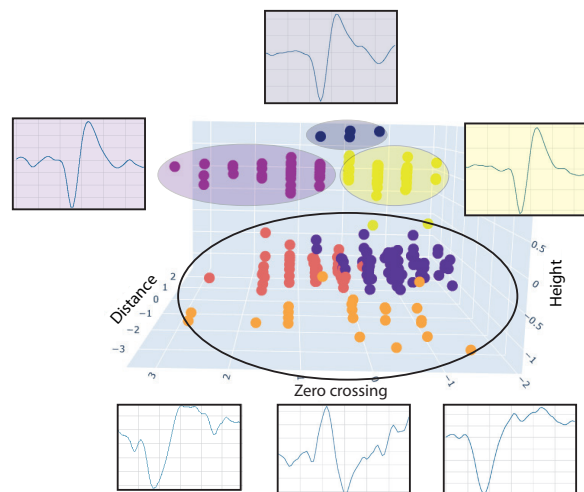
Figure 7. A 3D representation of the groups for the spikes in the monkey signal after normalization.

Five distinct groups emerge after applying K-means to the detected spikes within the human pancreatic signal. Among these, groups 1 and 4 yield the most favorable outcomes. Notably, in these groups, the spike shapes are distinctly defined, in contrast to the remaining groups, where residual noise in the spike shapes persists. This phenomenon arises due to the complexity of this particular signal, whose amplitude is in the nV range, which can clearly be mistaken for noise (see Figure A1b).

Upon observing the data distribution graphs in Figure 5c,d, it becomes evident that while distinct groups are identifiable, some clusters remain in close proximity to each other. This proximity raises the potential for confusion within the K-means algorithm, as certain clusters contain elements mixed with their neighboring counterparts. Consequently, spikes become challenging to differentiate, leading to their association with clusters, suppressing noise detection. To address this issue, Principal Component Analysis (PCA) was integrated for dimensionality reduction, aiming to enhance outcomes using the K-means method. The PCA method is described in the following section.

3.2. Principal Component Analysis (PCA) Method

Principal Component Analysis (PCA) [29] stands as a key technique in the field of signal analysis, especially when confronted with multi-dimensional datasets such as neural recordings. The fundamental goal of PCA is to reduce the complexity of high-dimensional data by transforming them into a lower-dimensional space while retaining the maximum variability present in the original data. In the context of spike sorting from biosignals, PCA serves as a powerful tool to extract the most informative features from the recorded waveforms. By identifying the principal components orthogonal axes along which data vary the most, PCA effectively reorients the data, allowing for a more intuitive representation that often captures the underlying patterns and relationships among different spikes. This dimensionality reduction not only aids in visualization but also facilitates subsequent analysis techniques, such as clustering algorithms, by reducing noise and improving the separability of distinct spike shapes. In this regard, PCA improves the precision and efficiency of spike-sorting procedures and provides more insightful interpretations drawn from neural recordings.

Figure 8 provides insight into the data distribution following dimensionality reduction to three principal components. The centroids of the groups obtained via K-means are prominently displayed. Furthermore, a 3D representation of the formed groups is depicted in Figure 9. Notably, this representation illustrates the segregation of three distinct spike groups for the human pancreatic signal instead of the two initially found by the K-means without PCA (Table 6), which struggled to achieve such differentiation. Moreover, PCA exhibits the capability to identify an additional group within the macaque monkey signal, a spike group that was initially misconstrued as noise by the K-means algorithm. This misclassification stemmed from the fact that the spike and the noise signal shared identical amplitudes as can be observed in Figure A5. The affirmation of a shape's correspondence to a spike depicted in Figure A5a stems from a distinct classification criterion. Specifically, a total of 30 samples were grouped in this particular spike configuration. This contrasts with the shape depicted in Figure A5b, where only a single signal was identified. This discrepancy hints at a potential distinction, the former shape aligning with a grouped spike, possibly originating from a nearby neuron, while the latter's sole detection might suggest a spike emitted by a more distant neuron [30]. The discernible difference in amplitude between the two shapes contributes to this analysis, corroborating their distinct neural origins. These results probe how PCA, by extracting salient features, contributes to the improved effectiveness of the K-means clustering method.
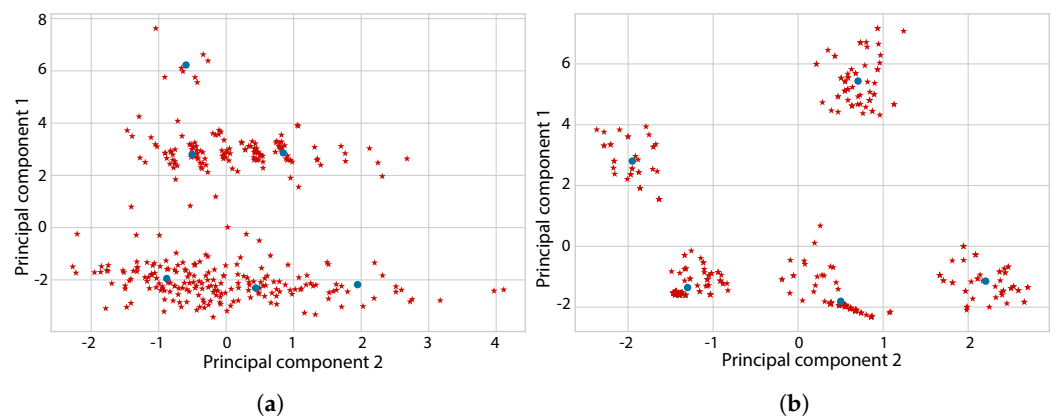
Figure 8. Distribution of clusters resulting from the application of K-means in conjunction with PCA on detected spikes from (**a**) macaque monkey signals and (**b**) human pancreatic signals.
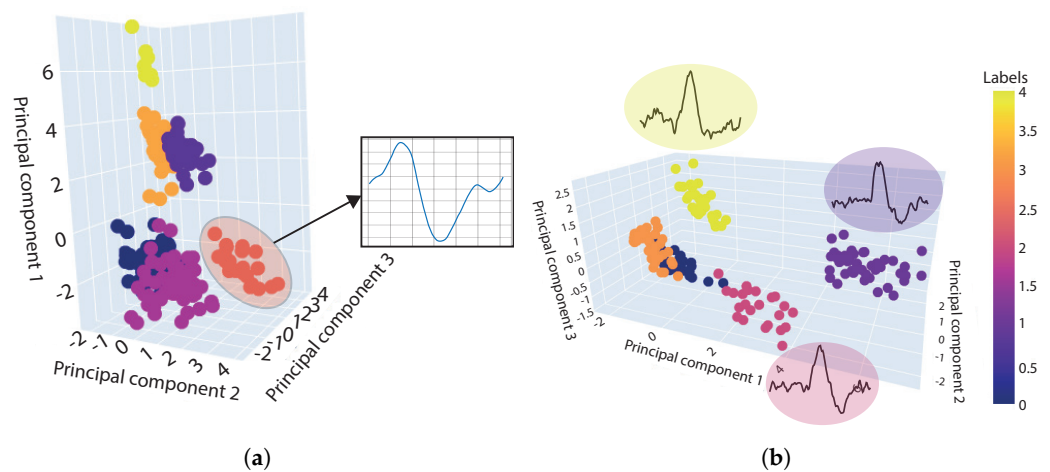

Figure 9. A 3D distribution graph of clusters resulting from the application of K-means in conjunction with PCA on newly detected spikes from (**a**) macaque monkey signals and (**b**) human pancreatic signals.

3.3. Hierarchical Clustering

Like the K-means method, hierarchical clustering is an unsupervised algorithm [31,32]. This technique constructs a series of nested clusters, forming a hierarchy typically visualized as a tree-like structure known as a dendrogram. This graph provides a visual representation of the cluster hierarchy. The hierarchical clustering process follows a bottom–up approach, where each data point initiates as an independent cluster, subsequently merging clusters based on a chosen metric (e.g., Euclidean distance). In other words, the algorithm treats each sample as an individual cluster and employs a defined merging strategy to combine clusters progressively. This iterative merging continues until the top of the tree is reached, culminating in forming the root cluster that encompasses all samples.

In the hierarchical clustering method, we start from the dendrogram that represents our data. In the K-means method, we observe that the normalized data present better results, allowing the algorithm to train in a more efficient way. In this case, the data used also correspond to those normalized; Figure 10 shows the dendrogram formed with the data of the feature matrix of normalized characteristics.

The dendrogram graph is a reference for determining the optimal number of clusters to create. A straightforward approach to selecting this number is employing a horizontal line as demonstrated in Figure A3. The count of vertical lines intersecting the horizontal line designates the number of clusters. While the figure distinctly illustrates two discernible groups, limiting the clusters to just two might result in the loss of valuable information. By analyzing the dendrogram's distribution, a strategic placement of the horizontal line

is established as exemplified in Figure 11. This choice yields a cluster count of six for both signals.
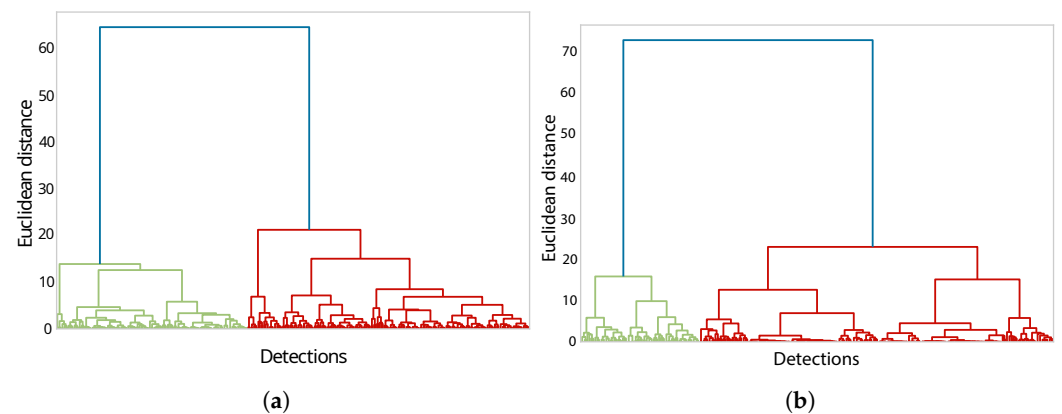
Figure 10. Dendrograms with normalized data from (**a**) macaque monkey signal and (**b**) human pancreatic signal. Where data detection at the bottom begins forming new groups based on distance similarity.
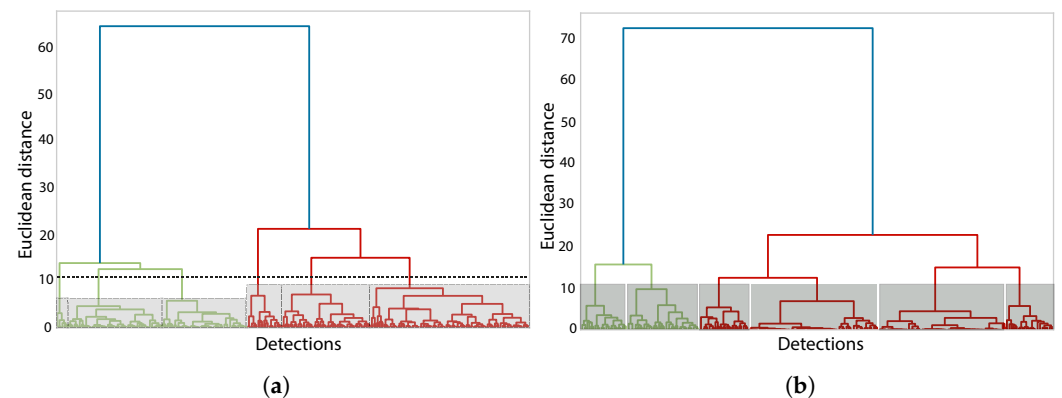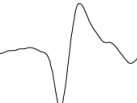

Figure 11. Selected clusters from the dendrograms for (**a**) macaque monkey data and (**b**) human pancreatic data. Where data detection at the bottom begins forming new groups based on distance similarity.

In the case of detection conducted on the macaque monkey signal, the resultant grouping closely resembles that derived from the K-means approach. The distribution of the 327 spikes across these groups is enumerated in Table 7. Evidently, it becomes possible to discern the three distinct types of potentials depicted in Figure 6, notably spanning groups 3, 4, and 5. Nonetheless, the K-means method supplemented with PCA more accurately distinguishes the distant potentials. It achieves this by effectively minimizing confusion with noise. However, it is worth noting that while the hierarchical grouping method endeavors to isolate the distant potentials within a group, it still exhibits a degree of blending with noise. Nevertheless, despite this challenge, the outcomes attained through the hierarchical grouping method slightly surpass those derived from K-means without PCA, signifying incremental improvement. Similar observations extend to the clusters formed from the spikes within the human pancreatic signal, mirroring the findings. The detail of these clusters can be observed in Table 8.

The data distribution and clusters for the spikes in the macaque monkey and human pancreatic signals are shown in Figure 12a,b.

Table 7. Groups formed with hierarchical clustering (macaque monkey data).
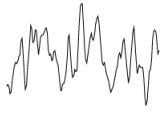
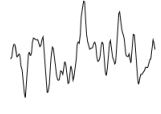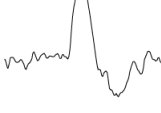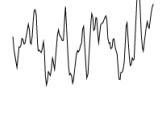| Group	Spike Shape	Number of Spikes
0		110
1		61
2		24
3		59
4		8
5		65

(**a**)   (**b**)
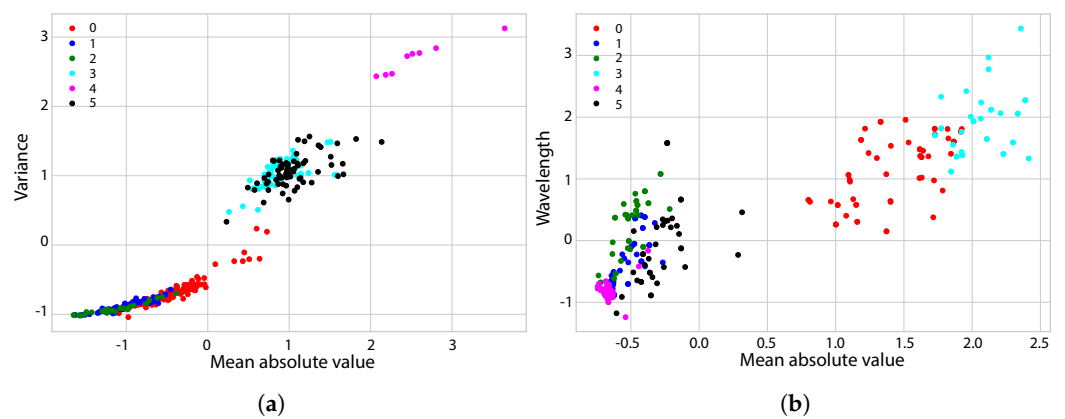
Figure 12. Hierarchical clustering from the detected spikes within the (**a**) macaque monkey and (**b**) pancreatic human data.

Table 8. Groups formed with hierarchical clustering (human pancreatic data).

Group	Spike Shape	Number of Spikes
0		59
1		106
2		39
3		38
4		104
5		40

3.4. Self-Organizing Maps (SOMs)

One of the major contributions of this work is the use of Self-Organizing Maps in the spike sorting of biosignals, as this type of neural network had not been used in this type of signal. A SOM, also known as a Kohonen map [33,34], is a type of artificial neural network used for unsupervised learning and the visualization of high-dimensional data. SOMs are particularly effective for clustering and dimensionality reduction tasks. The network consists of a grid of interconnected nodes (see Figure 13), each representing a specific region in the input data space. During training, the SOM learns to map input data points to nodes in a way that preserves the underlying data topology. Neighboring nodes on the grid respond similarly to similar input patterns, leading to a self-organized representation of the input data in a lower-dimensional space. This arrangement enables the visualization of complex data relationships and grouping of similar data points into clusters on the grid. SOMs are especially useful for exploratory data analysis, pattern recognition, and data visualization.

Mean absolute value
VarianceLongitud
Wavelength
RMS
Simple square integral
Integral of the absolute value
Height
Distance
Zero crossing

(**a**)

Mean absolute value Variance
Wavelength
RMS
Simple square integral
Integral of the absolute value
Height
Distance
Zero crossing

(**b**)
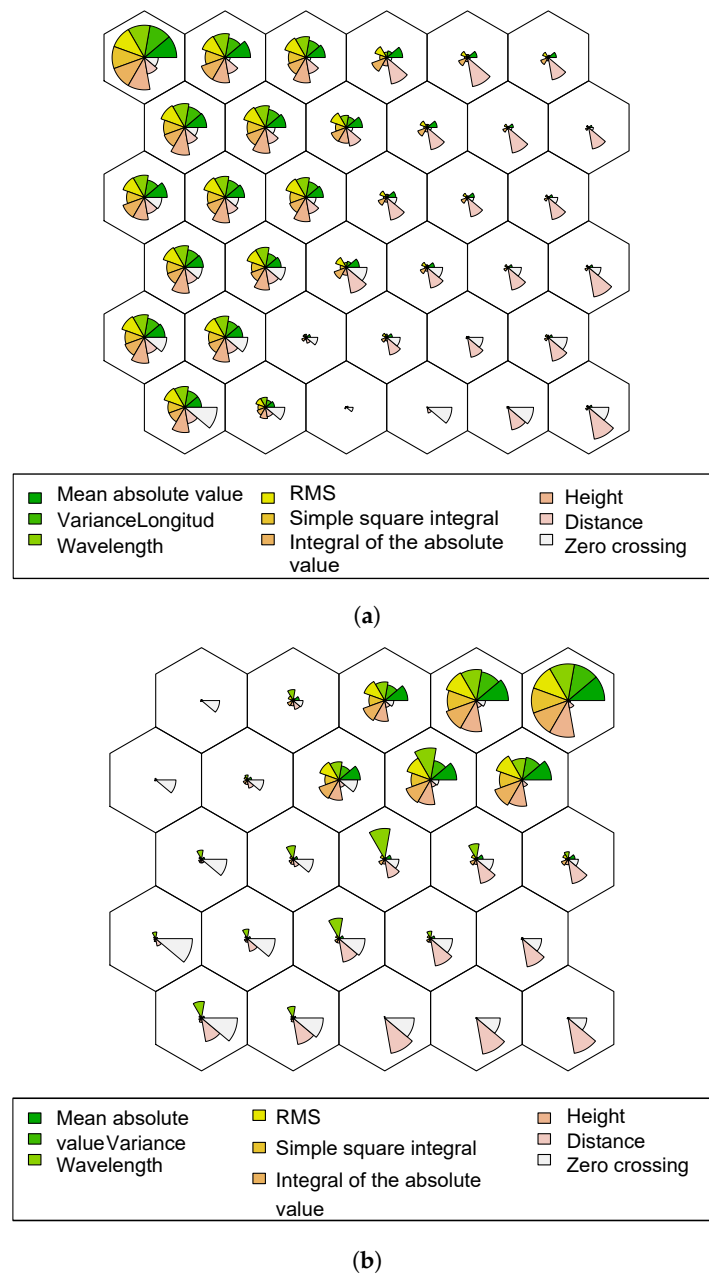
Figure 13. (**a**) A 6 × 6 map for the macaque monkey data and (**b**) a 5 × 5 map for the human pancreatic data.

Figure 13 provides insight into the weighting of each input variable across various segments of the map. This visualization allows us to discern the significance of each variable within each node of the map. However, a more comprehensive understanding of the variable importance within the map is encapsulated in Figure A4. In this illustration, a set of contrast graphs is presented for each individual variable. These visual representations offer a clearer indication of the variables that wield the most influence when activating each neuron within the map, thereby dictating the assignment of input data to these specific nodes.

Figure A4 shows the predominant impact of each variable across distinct regions of the map through the pronounced red and orange hues. Unlike the representation in Figure 13, this graph deliberately amplifies the contrast between areas with high and low values. This emphasis serves to facilitate a more intuitive grasp of the variable influence distribution within the map.

Figure 14a exhibits a density graph, an illustrative visualization that facilitates the identification of regions within the map housing the highest concentration of instances. This density mapping is depicted using a color scale that conveys the prevalence of data points within each node. It is important to note that the nodes within the map themselves do not directly correspond to the formed groups. Rather, these groups emerge later through the aggregation of nodes based on common characteristics.

To assemble the groups, the method involves calculating the Euclidean distance between each neuron and its neighboring nodes. Illustrated in Figure 14b, the distance graph offers a visual representation of the formed map, facilitating the subsequent formation of groups. This process entails aggregating neurons sharing a comparable distance, as nodes belonging to the same group typically exhibit proximity to each other.

(**a**)   (**b**)
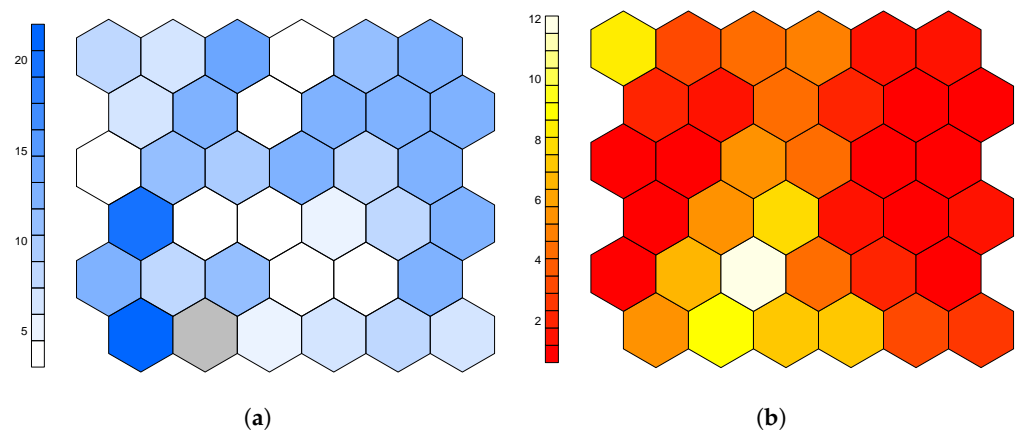
Figure 14. (**a**) Density and (**b**) distance graphs.

As previously highlighted, it is imperative to note that the nodes within the map do not inherently equate to groups. However, these nodes are structured in a manner where adjacent nodes share analogous weight vectors. Consequently, each map node operates as a precursor to a group and subsequently undergoes hierarchical grouping. The outcome of this process is evident in the dendrograms depicted in Figure A6. These dendrograms serve as a guiding framework to ascertain the optimal number of groups to form from the detected spikes for the macaque monkey (Figure A6a) and human pancreatic data (Figure A6b). Notably, through careful analysis, a decision was reached to shape six distinct groups for the former and five for the latest. This choice stems from the discernment of six groups effectively capturing differentiable disparities among the detected potentials as observed through the utilization of previous grouping methodologies. Moreover, aligning this determination with the distribution of the dendrogram attests to the judicious selection of the resulting groups.

With the optimal number of groups identified, the subsequent step involves establishing the grouping within the Self-Organizing Map. Illustrated in Figure 15, this graphical representation highlights the distinct clusters that emerge as a result of uniting nodes with the closest Euclidean distance. These clusters manifest as an embodiment of cohesive neural patterns, effectively grouping nodes that exhibit similar characteristics. Notably, Figure A7 offers a 3D visualization portraying the spatial distribution of these clusters, affording an insightful view into the intricate relationship among grouped neural activities.

Through this approach, we successfully delineate the three primary categories of potentials depicted in Figure 6 for the macaque monkey data. Furthermore, our results closely parallel those attained via the K-means method augmented with PCA in the accurate identification of distant potentials. The distinctive profiles characterizing the groups are laid out in Tables 9 and 10, providing an overview of the shape attributes intrinsic to each group for both signals, complemented by the corresponding number of constituent elements.
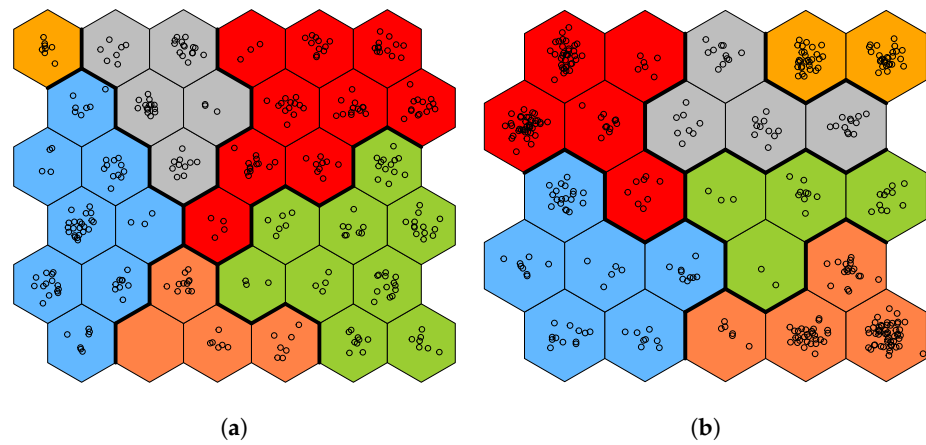
(**a**)   (**b**)

Figure 15. Groups formed on the Self-Organizing Map for (**a**) macaque monkey data and (**b**) human pancreatic data.

Table 9. Groups formed with Self-Organizing Maps (macaque monkey data).
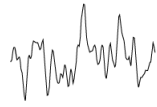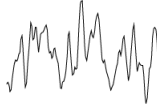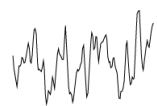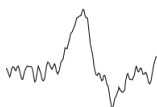
Group	Spike Shape	Number of Spikes
1		27
2		24
3		76
4		93
5		49
6		8

Table 10. Groups formed with Self-Organizing Maps (Human pancreatic data).

Group	Spike Shape	Number of Spikes
1		61
2		105
3		25
4		98
5		41
6		56

4. Discussion

The comprehensive exploration of clustering methods within the context of spike sorting in both macaque monkey and human pancreatic signals has yielded valuable insights into the intricacies of neural signal analysis. The amalgamation of advanced computational techniques with the distinct characteristics of neural recordings underscores the multi-faceted nature of this endeavor.

The initial step of spike detection formed the foundation for subsequent analysis. The application of an adaptive threshold, as described, facilitated the automatic identification of action potentials while efficiently adapting to background noise levels. Notably, this strategy allowed us to focus on the nuanced shapes and patterns of the detected spikes without being encumbered by noise artifacts.

Feature extraction emerged as an intermediate stage, enhancing the interpretability of the data while reducing their dimensionality. The utilization of Principal Component Analysis (PCA) for signal dimensionality reduction effectively captured essential information in both signals. This reduction in dimensionality proved beneficial for subsequent clustering algorithms, enabling more effective differentiation among distinct spike shapes. In the context of the macaque monkey signal, this approach enabled the identification of a cluster of spikes that were initially misinterpreted as noise due to their similar signal

amplitude. In the scenario of the human pancreatic signal, the combined utilization of PCA and K-means led to enhanced classification outcomes. Specifically, it transformed a scenario with only two distinct groups achieved through K-means alone into one featuring three well-defined groups, underscoring the complementary power of these methods when employed together.

The integration of Self-Organizing Maps (SOM) introduced a novel dimension to the study. The SOM methodology offered an unsupervised learning framework that revealed underlying data topology. The resulting clusters enabled the identification of similar neural patterns, thereby contributing to a refined understanding of neural activity. The delineation of clusters as showcased through density graphs and distance measurements provided a visual representation of the spatial distribution of these neural patterns.

The hierarchical clustering approach further refined the clustering process, allowing for the formation of cohesive groups based on the proximity of nodes. This hierarchical arrangement effectively organized the neural patterns into clusters that aligned with their inherent similarities. The discernment of optimal group numbers through dendrogram analysis attested to the efficacy of this approach in capturing meaningful clusters while avoiding over-segmentation. The results of this method are similar to those obtained jointly by PCA and K-means. This indicates that the intermediate use of PCA can be omitted if we focus on hierarchical clustering.

A comparison of the results achieved through different methods unveils noteworthy findings. The identification of the six types of potentials demonstrates the ability of the proposed approach to disentangle complex neural patterns. Furthermore, the resemblance of the results to the K-means method augmented with PCA in the identification of distant potentials underscores the robustness of the methodology.

5. Conclusions

This study presents a significant advancement in spike sorting by strategically combining established techniques with a pioneering application: Self-Organizing Maps (SOMs) for biosignal analysis. SOMs excel in unsupervised learning and dimensionality reduction, revealing the complexities of neural signal analysis from macaque monkey and human pancreatic signals. This multi-faceted endeavor unveils the nuanced effectiveness of SOMs and leverages them to form cohesive neural patterns through hierarchical clustering. Our approach successfully classified the neural data, achieving results comparable to established methods. Building on the foundation of adaptive thresholding, which effectively isolated action potentials while filtering background noise, PCA emerged as a crucial step for dimensionality reduction. This enhanced data interpretability and facilitated the identification of previously misclassified noise spikes in macaque monkey signals. The combined power of PCA and K-means in human pancreatic signals transformed a binary classification into one with three distinct groups. The introduction of SOMs added a novel unsupervised learning dimension. SOMs delineated clusters based on spatial distribution by revealing underlying data topology, offering a refined perspective on neural signal organization. Hierarchical clustering further refined this process, forming cohesive groups based on node proximity. This method effectively captured significant clusters without over-segmentation, achieving results comparable to PCA and K-means, suggesting PCA might be optional for specific analyses.The proposed approach successfully disentangled complex neural patterns, identifying six types of potentials. This robustness, mirrored in the similarity of results to the PCA-augmented K-means method, underscores the efficacy of the methodologies. Future research can explore additional clustering algorithms, feature extraction methods, and more sophisticated noise-filtering techniques. Extending this analysis to various real-world scenarios where other types of biosignals could validate the generalizability of these methods, this will be imperative for future research. Additionally, investigating real-time implementations and integrating these techniques into clinical applications or brain–computer interfaces is essential for improving signal processing and

information decoding for real-world applications, thus offering promising avenues for advancing neural science and its practical applications.
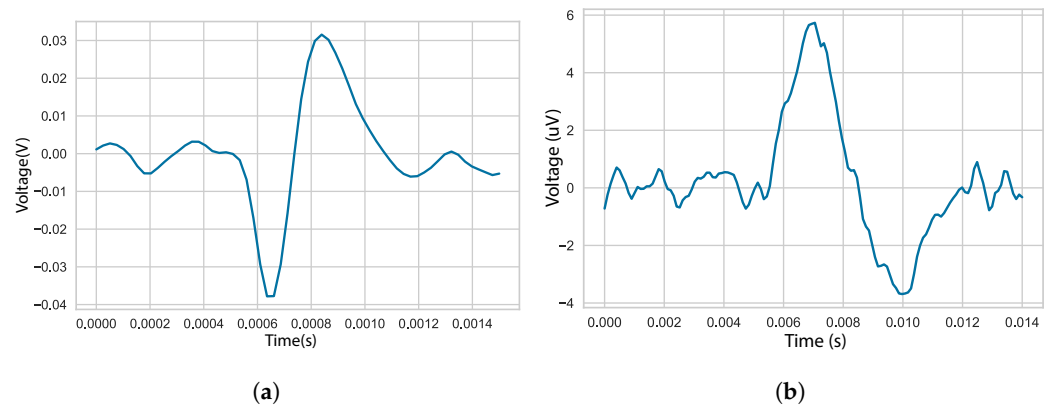
Appendix A

Figure A1. Spike shapes found within the (**a**) macaque monkey signal, (**b**) human pancreatic signal.
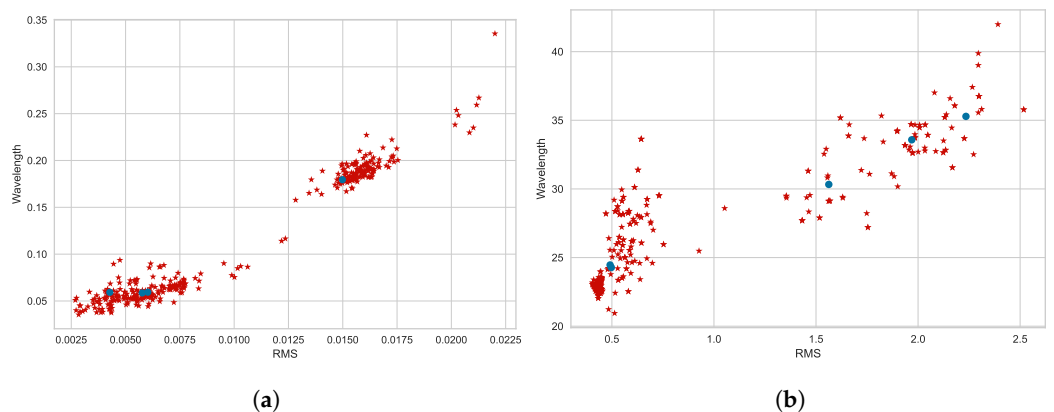

Figure A2. K-means clustering for spikes in both signals (**a**) macaque monkey and (**b**) human pancreatic.
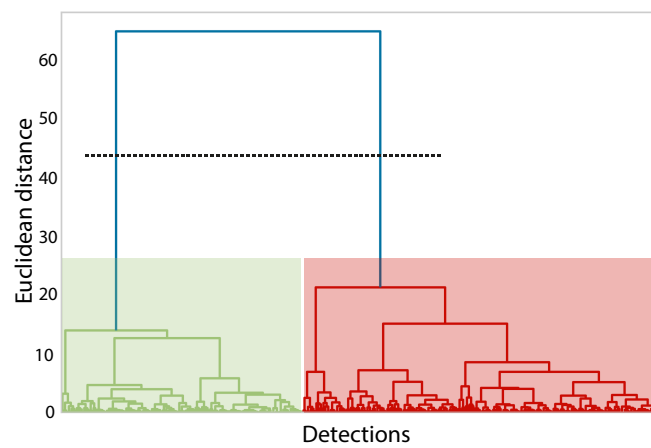
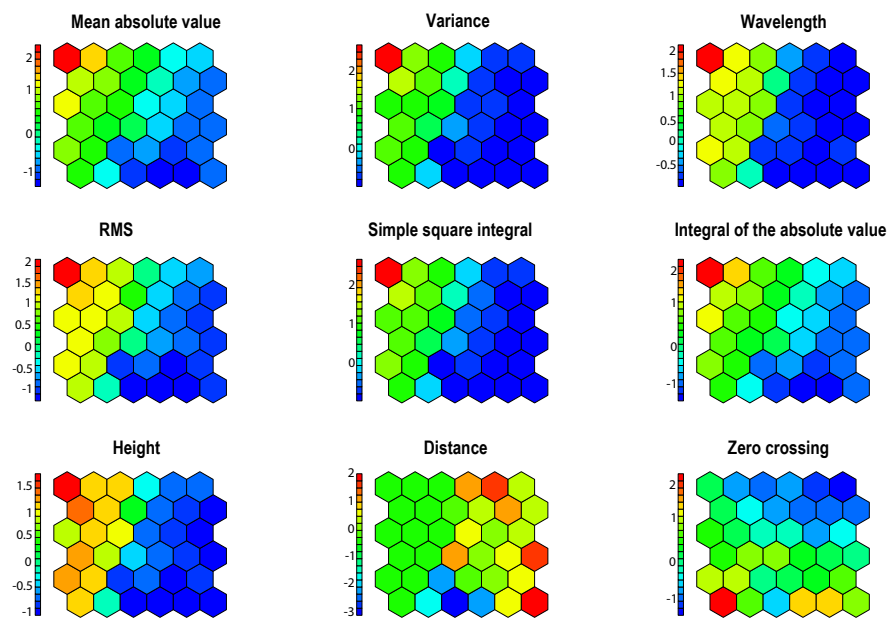**Figure A3.** Example of the selection of the number of groups in the dendrogram of the macaque monkey data.

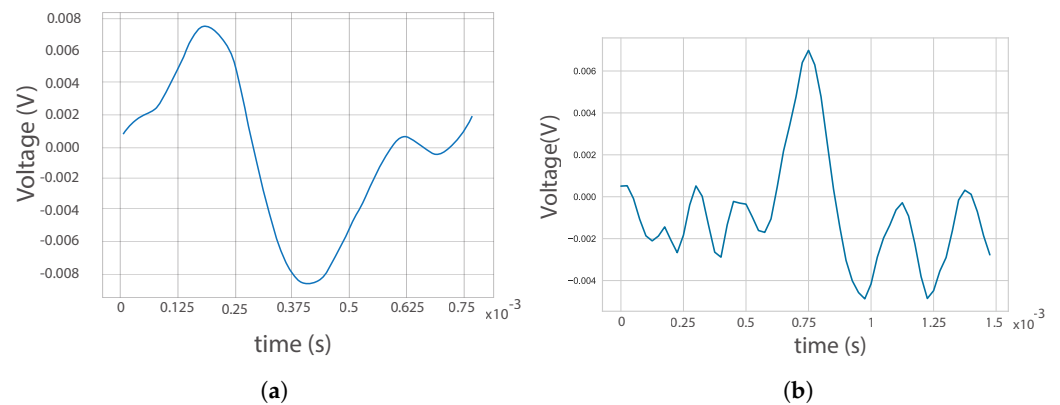

Figure A4. Color map of variables.

(**a**)	(**b**)

Figure A5. (**a**) Spike shape of the new group detected with K-means and PCA from the detected spikes in the macaque monkey signal. (**b**) Noise signal with the same amplitude as the spike.
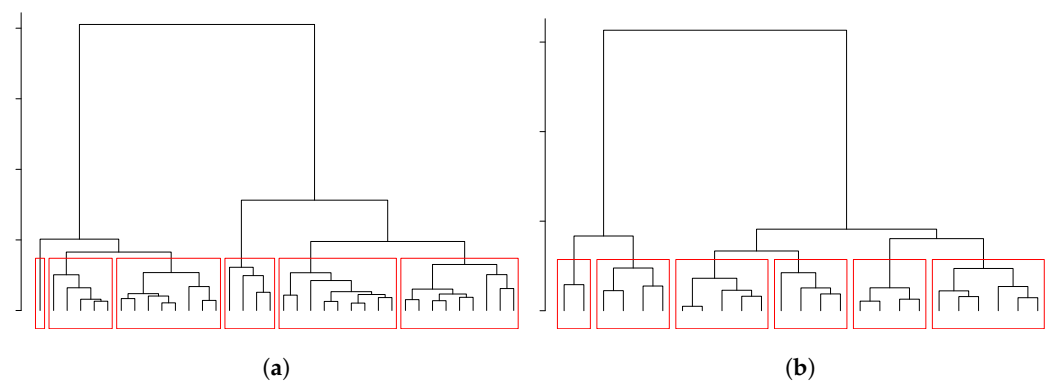
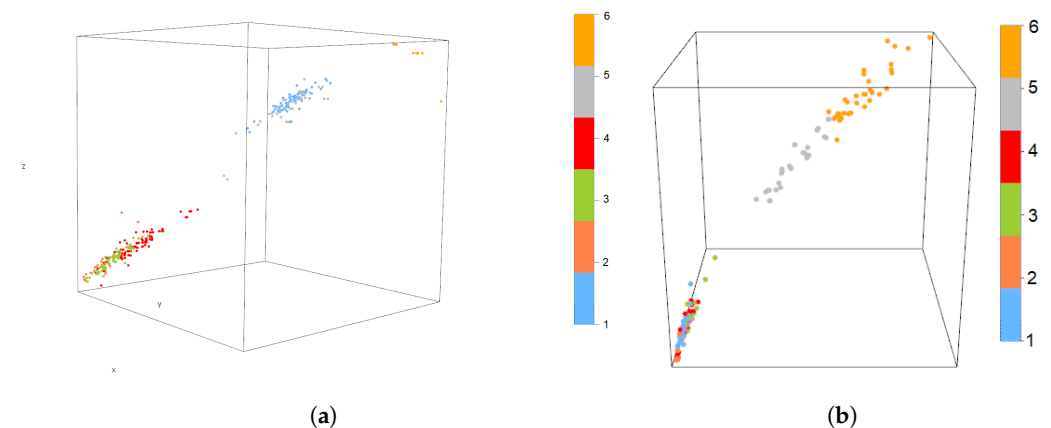**Figure A6.** Dendrogram of the map nodes for (**a**) macaque monkey data and (**b**) human pancreatic data.

Figure A7. Groups distribution for (**a**) macaque monkey data and (**b**) human pancreatic data.

References

1. Alber, M.; Buganza Tepole, A.; Cannon, W.R.; De, S.; Dura-Bernal, S.; Garikipati, K.; Karniadakis, G.; Lytton, W.W.; Perdikaris, P.; Petzold, L.; et al. Integrating machine learning and multiscale modeling-perspectives, challenges, and opportunities in the biological, biomedical, and behavioral sciences. *npj Digit. Med.* **2019**, *2*, 115. [CrossRef] [PubMed]
2. Lu, H.Y.; Lorenc, E.S.; Zhu, H.; Kilmarx, J.; Sulzer, J.; Xie, C.; Tobler, P.N.; Watrous, A.J.; Orsborn, A.L.; Lewis-Peacock, J.; et al. Multi-scale neural decoding and analysis. *J. Neural Eng.* **2021**, *18*, 045013. [CrossRef]
3. Wang, C.; Pesaran, B.; Shanechi, M.M. Modeling multiscale causal interactions between spiking and field potential signals during behavior. *J. Neural Eng.* **2022**, *19*, 026001. [CrossRef] [PubMed]
4. Lewicki, M.S. A review of methods for spike sorting: The detection and classification of neural action potentials. *Network* **1998**, *94*, R53–R78. [CrossRef]
5. Oweiss, K.; Aghagolzadeh, M. Chapter 2—Detection and Classification of Extracellular Action Potential Recordings. In *Statistical Signal Processing for Neuroscience and Neurotechnology*; Oweiss, K.G., Ed.; Academic Press: Oxford, UK, 2010; pp. 15–74. [CrossRef]
6. Buccino, A.P.; Garcia, S.; Yger, P. Spike sorting: New trends and challenges of the era of high-density probes. *Prog. Biomed. Eng.* **2022**, *4*, 022005. [CrossRef]
7. Urai, A.E.; Doiron, B.; Leifer, A.M.; Churchland, A.K. Large-scale neural recordings call for new insights to link brain and behavior. *Nat. Neurosci.* **2022**, *25*, 11–19. [CrossRef] [PubMed]
8. Caro-Martín, C.R.; Delgado-García, J.M.; 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**, *8*, 17796. [CrossRef] [PubMed]
9. Vogt, N. Benchmarked spike sorting. *Nat. Methods* **2020**, *17*, 656. [CrossRef]
10. Valencia, D.; Mercier, P.P.; Alimohammad, A. In vivo neural spike detection with adaptive noise estimation. *J. Neural Eng.* **2022**, *19*, 046018. [CrossRef]
11. Wilson, S.B.; Emerson, R. Spike detection: A review and comparison of algorithms. *Clin. Neurophysiol.* **2002**, *113*, 1873–1881. [CrossRef]
12. Rey, H.G.; Pedreira, C.; Quian Quiroga, R. Past, present and future of spike sorting techniques. *Brain Res. Bull.* **2015**, *119*, 106–117. [CrossRef] [PubMed]
13. Meyer, L.M.; Samann, F.; Schanze, T. DualSort: Online spike sorting with a running neural network. *J. Neural Eng.* **2023**, *20*, 056031. [CrossRef] [PubMed]

14. Saif-Ur-Rehman, M.; Ali, O.; Dyck, S.; Lienkämper, R.; Metzler, M.; Parpaley, Y.; Wellmer, J.; Liu, C.; Lee, B.; Kellis, S.; et al. SpikeDeep-classifier: A deep-learning based fully automatic offline spike sorting algorithm. *J. Neural Eng.* **2021**, *18*, 016009. [CrossRef] [PubMed]

15. Cam, S.L.; Jurczynski, P.; Jonas, J.; Koessler, L.; Colnat-Coulbois, S.; Ranta, R. A Bayesian approach for simultaneous spike/LFP separation and spike sorting. *J. Neural Eng.* **2023**, *20*, 026027. [CrossRef]

16. Li, Z.; Wang, Y.; Zhang, N.; Li, X. An Accurate and Robust Method for Spike Sorting Based on Convolutional Neural Networks. *Brain Sci.* **2020**, *10*, 835. [CrossRef] [PubMed]

17. Okreghe, C.; Zamani, M.; Demosthenous, A. A Deep Neural Network-Based Spike Sorting with Improved Channel Selection and Artefact Removal. *IEEE Access* **2023**, *11*, 15131–15143. [CrossRef]

18. Ardelean, E.R.; Coporîie, A.; Ichim, A.M.; Dînșoreanu, M.; Mureşan, R.C. A study of autoencoders as a feature extraction technique for spike sorting. *PLoS ONE* **2023**, *18*, e0282810. [CrossRef] [PubMed]

19. Wang, M.; Zhang, L.; Yu, H.; Chen, S.; Zhang, X.; Zhang, Y.; Gao, D. A deep learning network based on CNN and sliding window LSTM for spike sorting. *Comput. Biol. Med.* **2023**, *159*, 106879. [CrossRef]

20. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]

21. Li, S.; Tang, Z.; Yang, L.; Li, M.; Shang, Z. Application of deep reinforcement learning for spike sorting under multi-class imbalance. *Comput. Biol. Med.* **2023**, *164*, 107253. [CrossRef]

22. Saif-Ur-Rehman, M.; Lienkämper, R.; Parpaley, Y.; Wellmer, J.; Liu, C.; Lee, B.; Kellis, S.; Andersen, R.; Iossifidis, I.; Glasmachers, T.; et al. SpikeDeeptector: A deep-learning based method for detection of neural spiking activity. *J. Neural Eng.* **2019**, *16*, 056003. [CrossRef] [PubMed]

23. Soleymankhani, A.; Shalchyan, V. A New Spike Sorting Algorithm Based on Continuous Wavelet Transform and Investigating Its Effect on Improving Neural Decoding Accuracy. *Neuroscience* **2021**, *468*, 139–148. [CrossRef] [PubMed]

24. Zhang, T.; Azghadi, M.R.; Lammie, C.; Amirsoleimani, A.; Genov, R. Spike sorting algorithms and their efficient hardware implementation: A comprehensive survey. *J. Neural Eng.* **2023**, *20*, 021001. [CrossRef]

25. Iniguez-Lomeli, F.J.; Bornat, Y.; Renaud, S.; Barron-Zambrano, J.H.; Rostro-Gonzalez, H. A real-time FPGA-based implementation for detection and sorting of bio-signals. *Neural Comput. Appl.* **2021**, *33*, 12121–12140. [CrossRef]

26. Harrison, R. A low-power integrated circuit for adaptive detection of action potentials in noisy signals. In Proceedings of the 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (IEEE Cat. No. 03CH37439), Cancun, Mexico, 5 April 2004; Volume 4, pp. 3325–3328. [CrossRef]

27. Lloyd, S. Least squares quantization in PCM. *IEEE Trans. Inf. Theory* **1982**, *28*, 129–137. [CrossRef]

28. Kanungo, T.; Mount, D.; Netanyahu, N.; Piatko, C.; Silverman, R.; Wu, A. An efficient k-means clustering algorithm: Analysis and implementation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 881–892. [CrossRef]

29. Greenacre, M.; Groenen, P.J.F.; Hastie, T.; D'Enza, A.I.; Markos, A.; Tuzhilina, E. Principal component analysis. *Nat. Rev. Methods Prim.* **2022**, *2*, 100. [CrossRef]

30. Pedreira, C.; Martinez, J.; Ison, M.J.; Quian Quiroga, R. How many neurons can we see with current spike sorting algorithms? *J. Neurosci. Methods* **2012**, *211*, 58–65. [CrossRef] [PubMed]

31. Shahid, N. Comparison of hierarchical clustering and neural network clustering: An analysis on precision dominance. *Sci. Rep.* **2023**, *13*, 5661. [CrossRef]

32. Cabezas, L.M.; Izbicki, R.; Stern, R.B. Hierarchical clustering: Visualization, feature importance and model selection. *Appl. Soft Comput.* **2023**, *141*, 110303. [CrossRef]

33. Kohonen, T. Self-organized formation of topologically correct feature maps. *Biol. Cybern.* **1982**, *43*, 59–69. [CrossRef]

34. Kohonen, T. *Self-Organizing Maps*, 3rd ed.; Springer Series in Information Sciences; Springer: Berlin, Germany, 2000.