# Special Session 1: Neural Decoding and Brain Computer Interface:

# Artificial Neural Network-assisted Amplitude Thresholding Improves Spike Detection

XIANG CHENG

Research Center for Brain-inspired Intelligence, Institute of Automation, Chinese Academy of Sciences; chengxiang2018@ia.ac.cn

XUAN HAN

Beijing University of Technology, Research Center for Brain-inspired Intelligence, Institute of Automation, Chinese Academy of Sciences; hanxuanxuan@emails.bjut.edu.cn

YU SONG

Research Center for Brain-inspired Intelligence, Institute of Automation, Chinese Academy of Sciences; songyu18@mails.ucas.ac.cn

TIELIN ZHANG

Research Center for Brain-inspired Intelligence, Institute of Automation, Chinese Academy of Sciences; tielin.zhang@ia.ac.cn

BO XU

Research Center for Brain-inspired Intelligence, Institute of Automation, Chinese Academy of Sciences; xubo@ia.ac.cn

As brain-related research presents increasing importance, the requirement for automatic spike detection algorithms also emerges. Traditional spike detection algorithms, including amplitude thresholding and wavelet transformation, show several shortcomings that impede the practical application. Here, we propose an artificial neural network-assisted amplitude thresholding algorithm and conduct experiments with raw signals collected from the primary somatosensory cortex and primary motor cortex of macaques. Using F1 score as an evaluation index, artificial neural networks, as well as its lightweight version, effectively help the amplitude thresholding to achieve better performance, showing enormous potential for real-time spike detection application.

**CCS CONCEPTS** • Computing methodologies~Machine learning~Learning paradigms~Supervised learning~Supervised learning by classification • Applied computing~Life and medical sciences~Bioinformatics

**Additional Keywords and Phrases:** Spike detection, Amplitude thresholding, Artificial neural network

# 1 INTRODUCTION

Nowadays, research related to the brain is presenting increasing importance and driving the requirements for specific inspection equipment, which records the activity trace of neurons that contribute to brain functions. Spike detection[1, 2], as a previous step of spiking sorting, shows great importance to extracellular single-unit recording in brain-computer interface (BCI)[3], diagnoses of nervous system diseases[4], and research on brain functions. By picking up the timing of action potentials in continuous input signals, raw waveform data will be transformed into spike trains representing the activity patterns of recorded neurons. Due to the large volume of raw signals with multiple channels and high sampling rates, applying automatic algorithms to replace manual data processing is essential for the practical application of spike detection.

However, there are still difficulties remaining for automatic spike detection algorithms[2]. On the one hand, filtrating spike signals from raw extracellular neural data is challenging, as raw data are vastly influenced by different neuron types, as well as the positional relationship among target neurons, sampling electrodes, and other nontarget neurons, leading to the diversity of waveform in spike signals. Additionally, the noise signals generated from irrelevant neural activities, sampling electrode itself, and interference among other adjacent electrodes may possess similar features to spike signals, including waveform, amplitude, and width, thus even confusing seasoned neuroscientists. On the other hand, computational resources in the practical application may be limited in many aspects such as the size and power consumption of the hardware[5]. Considering an invasive BCI scenario where the spike detectors need to be miniaturized and the influence of the heat loss on brain physiological activities should be minimized, the dilemma between performance and lightweight brings more problems to the design of spike detection algorithms.

Despite all the difficulties, related researchers have proposed several spike detection algorithms. Amplitude thresholding for one or both edges is often used in the real-time detection of BCI systems for its low computational costs[6, 7]. However, this method depends on unreliable priori assumptions of noise amplitude and can hardly handle the situation when the noise level is close to or surpasses the amplitude peak of action potentials. Template matching achieves better performance than simple thresholding[8], but shows more dependence on a priori knowledge about the signal shapes and detection algorithm for template generating[9], leading to similar issues as those in simple thresholding and additional algorithm complexity. Nonlinear energy operator (NEO) is another widely used threshold-based method that amplifies the amplitude and frequency of local peaks and also requires manual or automatic adjustment[10, 11]. Wavelet transformation-based algorithms (WT) are another major category for spike detection whose efficiency of spike largely depends on the selection of the mother wavelet[12,13].

In this work, we choose amplitude thresholding with self-adaption of threshold as the baseline for its simplicity and low energy consumption on real-time spike detection and test its performance on neural signals collected from the primary somatosensory cortex and primary motor cortex of macaques. Furtherly, we apply an artificial neural network (ANN) as a pattern recognizer to judge the detected signals once again after the regular spike detection module and also consider the lightweight of the neural network. Experimental results show that the introduction of ANNs and lightweight ANNs effectively improve the performance, compared to those of simple self-adaptive amplitude thresholding.

## 2 METHODS

### 2.1 Data collecting

Raw neural data we use in this work is obtained by multi-shank nanoelectronic thread (NET)[14] probes which are implanted into the primary somatosensory cortex and primary motor cortex of macaques. The probe tip depth ranges from 700 μm to 1mm. Neural signals are collected using the Intan 128-channel RHD2164 evaluation system (Intan Technology) at a sampling rate of 30 kHz. The commonly used median reference is applied to the raw data to reduce modal noise. In vivo, impedance is measured at 1khz with the same system before each recording. Individual units are identified based on the criterion that the proportion of pulses with a pulse interval of less than 2 ms is less than 1% of the total discharge. We select one channel with a relatively high signal noise ratio and an average amplitude of 30μv from the records of 128-channel NET probes for the following experiments.
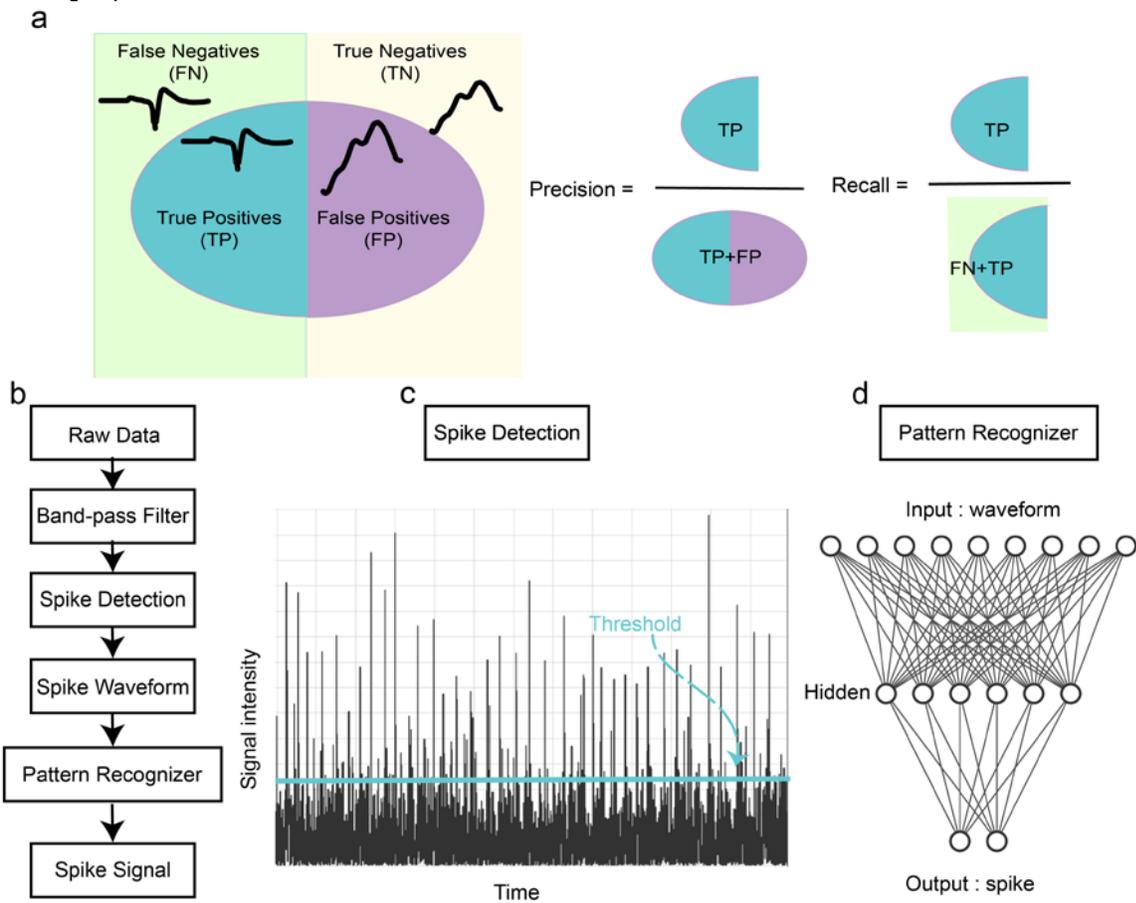


Figure 1: Diagram depicts the calculating of precision and recall and overview of spike detection algorithm.

## 2.2 Evaluation index

In this work, we assume the result of spike sorting performed by Offline Sorter (Plexon)[15] as ground truth. The spike signals detected by our algorithms are compared with the aforementioned ground truth to judge whether they belong to true positive samples or false positive samples, while those spike signals that exist in the ground truth but are not detected by our algorithm are defined as false negative samples. However, it is hard to define true negative samples in this task, as there is no spike taking place during most time of the sampling process and numerous true negative samples can be found, far more than the other three categories. Thus, the commonly used receiver operator characteristic (ROC) curve which needs to collect the number of true negative samples is unsuitable for this work and we choose F1 score which is the harmonic mean of the precision and recall to represent the performance of different methods. The detailed process of calculating precision, recall, and F1 score is shown in Fig. 1.a and as follows:

$$Precision = \frac{true\ positive}{true\ positive + false\ positive} \quad (1)$$

$$Recall = \frac{true\ positive}{true\ positive + false\ negative} \quad (2)$$

$$F_1 = \frac{2 precision \cdot recall}{precision + recall} \quad (3)$$

## 2.3 Overview of spike detection algorithm

In this work, we priorly follow a commonly used procedure of amplitude thresholding as shown in Fig. 1.b. The original signals are firstly preprocessed by a band-pass filter to eliminate noise and direct current offset in the original signal at 300-3000 Hz. Then an amplitude threshold (as shown in Fig. 1.c) is set to detect the after-depolarization peak membrane potential from the filtered signal.

To calculate the threshold, the NEO（as shown in Equation (4)) is firstly used to process the filtered data $X(t)$ at time $t$ and amplify the energy of the mutation signal to obtain better results in spike extraction. Then the mean absolute value (shorted as MAV, the same hereinafter) of data is calculated in one second as the unit of threshold value in the next second so that the threshold could be adjusted self-adaptively to the continuous input data. Additionally, the aforementioned MAV calculating only considers input information in a short period and neglects former historical information. Thus, we apply an iterative model (Equation (5)) in threshold calculation to retain the historical information and make the unit of threshold value more stable.

$$NEO: X(t) = X(t)^2 - X(t+1)(t-1)(4)$$

$$threshold\ unit(t) = \alpha \cdot mav(t-1) + (1-\alpha)mav(t) \quad (5)$$

After threshold selection followed by a normalization process, a whole spike waveform comprised of membrane potential at 34 neighboring sampling points is extracted according to the relative timing of the peak in action potential. Additionally, we introduce fully-connected ANNs (as shown in Fig. 1.d) to fulfill more accurate detection, which further estimates whether the normalized waveforms are real spike signals or above-threshold noise signals. As ANNs bring extra computational costs, especially in hardware implementation, we also consider several methods to lightweight the networks. ANNs and their lightweight methods are furtherly discussed in Section 2.4 while experiment results are shown in Section 3.2.

## 2.4 Artificial neural network and lightweight methods

ANNs are computational models[16] which are widely utilized on variant tasks, such as image classification[17], speech recognition[18, 19], and natural language processing[20, 21]. Inspired by biological

neural networks[22, 23], ANNs are constructed from artificial neurons which serve as computing nodes and imitate the information integration function of real biological neurons. After receiving input $x$ from upstream neurons with connection strength $w$ and bias $b$, artificial neurons convert integrated information with a non-linear function $\sigma(\cdot)$ into output $y$ and deliver the output to downstream neurons, as shown in Equation (6):

$$y = \sigma(wx + b) \quad (6)$$

As for network topology, we adopt a fully-connected structure where the network is divided into several sequential layers and neurons connect with all neurons from previous layer and next layer[24]. The first layer of ANNs, also called the input layer, has the same scale as the input data to be processed, which is the entire or part of the spike waveform obtained by traditional spike detection methods. The last layer of ANNs, also called the output layer, contains two neurons that represent positive and negative, respectively. If the output of the neuron representing positive surpasses that of the neuron representing negative, the input waveform is furtherly confirmed to be real spike signals. Otherwise, it is supposed that negative noise signals have been misjudged as positive spike signals by the former module. To ensure the classification accuracy of ANNs, a dataset consisting of manually selected positive and negative samples is used to adjust the connection weight and bias in ANNs with backpropagation algorithm.

Although the scale of ANNs in this work is relatively small, compared to nowadays deep artificial neural networks (DNNs), the requirement of lightweight design is still essential for specific application scenes, e.g. brain-computer interface (BCI), behavioral experiments. In this work, we test three approaches for lightweight. The first one is decreasing the number of hidden layers and an extreme example is a network containing only input and output layers that directly connect with each other. The second one is truncating the head and/or tail of the input waveform to reduce the dimension of input layer while conserving sufficient pattern information. The third one is weight quantization which is commonly used before the hardware implementation of ANNs. When initialized and trained on a mainstream deep learning framework, the connection weights and biases are floating point arithmetic but low-bit integers are needed. Among several weight quantization methods, we choose a relatively simple procedure where floating point arithmetic weights are first normalized, then multiplied by the maximum, and finally, the floor of the outcomes is adopted as the quantized weights.

## 3  EXPERIMENTS

### 3.1  Amplitude thresholding with self-adaptive threshold

We used multiples of the unit value (from twice to seven times) as threshold intensities. As threshold intensity was strengthened, the precision increased from 24.49% to 79.90% while recall decreased from 93.45% to 33.23% (shown in Fig. 2.a, b). As a result, the highest F1 score (0.59) was achieved when the threshold intensity was four times the unit value with moderate precision (53.47%) and recall (65.44%) (shown in Fig. 2.c). Two trials where the iterative model mentioned in Section 2.3 were applied ($\alpha$=0.99 and $\alpha$=0.9) were also presented in Fig. 2.d. However, the introduction of the iterative model showed no benefit but reduction in both recall and precision, which might be due to the instability of spike patterns along the continuous input data.
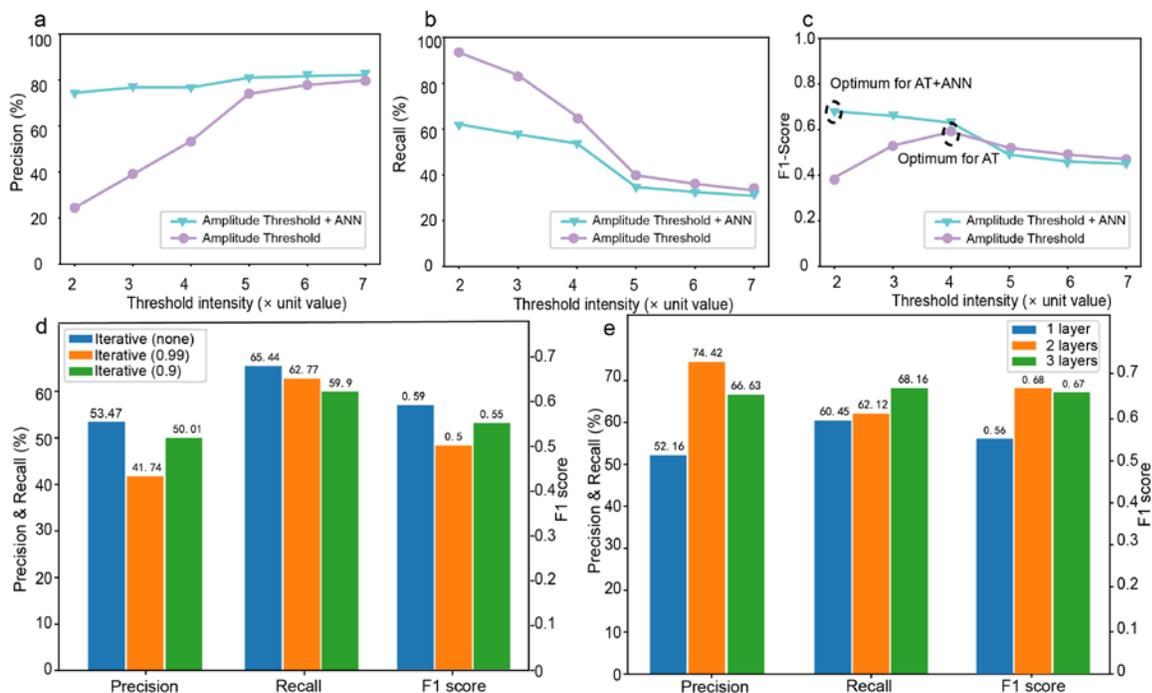
Figure 2: (a-c). Precision, recall, and F1 score of simple amplitude thresholding and ANN-assisted amplitude thresholding at different threshold intensities. Black dashed circles represent the highest F1 scores. d. Precision, recall, and F1 score of simple amplitude thresholding with no iterative models, iterative model with $\alpha$ =0.99 and $\alpha$ =0.9. e. Precision, recall, and F1 score of ANN-assisted amplitude thresholding with different layers of ANN.

## 3.2 Performances of ANN-assisted spike detection

Experimental results in Section 3.1 indicated that simple amplitude thresholding could hardly balance the precision and recall and thus achieved a relatively lower F1 score. As mentioned in Section 2.4, we introduced an ANN to furtherly judge the outcomes of the regular spike detection module. This additional mode discriminator was supposed to cut down a large number of false positive samples while discarding a small proportion of true positive samples, resulting in a considerable increase in precision and a relatively minor reduction in recall. When the spike detection module was at a threshold possessing high recall and low precision, the introduction of ANN would finally rise the F1 score. As shown in Fig. 2(a-c), when the threshold intensity was twice the unit value, the introduction of ANN raised precision from 24.49% to 74.42%. Though recall fell from 93.45% to 62.12%, the F1 score increased from 0.39 to 0.68, making this threshold intensity the best among the six intensities we tested. On the contrary, when the threshold intensity was seven times the unit value where the precision was relatively high but the recall was relatively low, the introduction of ANN damaged performance as recall fell from 33.23% to 30.80% and F1 score fell from 0.47 to 0.45, though precision increased from 79.90% to 82.22%.

Additionally, we also tested the potential performance loss of lightweight ANN which was necessary for practical applications in resource-limited scenarios like real-time invasive BCI. Three lightweight strategies were discussed.
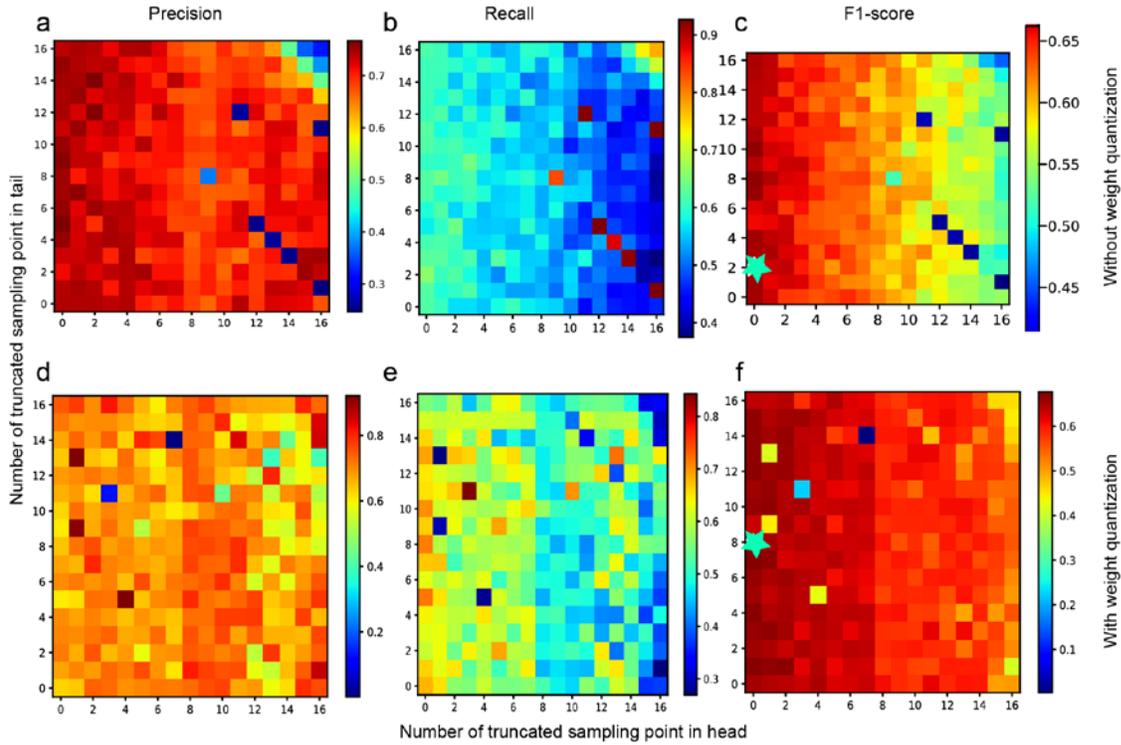
Figure 3: (a-c). Heat maps of precision, recall, and F1 score in trials without weight quantization. (d-f). Heat maps of precision, recall, and F1 score in trials with weight quantization. The cyan stars represent the highest F1 scores among other trials.

### 3.2.1 Different Numbers of Hidden Layers.

As shown in Fig. 2.e, among tested ANNs-assisted algorithms with different numbers of hidden layers with twice the unit value as threshold intensity, ANN with one hidden layer achieved better performance (F1 score: 0.68; precision: 74.42%; recall 62.12%). The reduction in the precision of ANN with two hidden layers (66.63%) might due to the overfitting problem in deeper ANNs.

### 3.2.2 Input Truncation.

As shown in Fig. 2(a-c), we applied heat maps to demonstrate the performances of the ANN-assisted algorithm in trails where different lengths of head and/or tails of input waveform were truncated. The longitudinal distribution of evaluation indices in Fig. 2(a-c) showed that the completeness of the fore part of the waveform presented more importance than the rear part. The conclusion was also proved by the fact that when two sampling points from the tail were truncated and zero sampling point from the head was truncated, the algorithms achieved the highest F1 score (0.69) among all contrasts.

### 3.2.3 Weight Quantization.

To test the influence of reduction in bit number on the performance of ANN-assisted algorithm, we transformed connection weights and biases from 32-bit signed floating point arithmetics to 9-bit signed integers and 8-bit unsigned integers, respectively[25]. The decrease in top performance (from 0.69 to 0.68) brought by weight quantization was minor, as shown in Fig. 2(d-f).

## 4 DISCUSSION

In this work, we test spike detection algorithms on collected neural signals and verify their limitation on balancing precision and recall. By introducing ANN, the dilemma is preliminarily mitigated. But there is still plenty of room for improvement. On the one hand, ANNs are data-driven computational models, and the quality of training datasets vastly influences the accuracy of ANN models. However, generally accepted standard training datasets have not been founded in spike detection tasks, and the manually selected training dataset we use in this work is far from perfect, implying a further improvement in the performance of ANN-assisted spike detection. On the other hand, as lightweight ANNs could improve the performance of an existing spike detection algorithm indicating the capability to extract pattern characteristics of action potentials from neural signals, an end-to-end ANN that takes raw signals within a time window as input and judges whether the input belongs to spike signals or not may have the potential to fulfill spike detection algorithm balancing high performance and low consumption.

## REFERENCES

[1] S. Kim and J. McNames, "Automatic spike detection based on adaptive template matching for extracellular neural recordings," *J. Neurosci. Methods*, vol. 165, no. 2, pp. 165–174, 2007.

[2] S. B. Wilson and R. Emerson, "Spike detection: a review and comparison of algorithms," *Clin. Neurophysiol.*, vol. 113, no. 12, pp. 1873–1881, 2002.

[3] J. R. Wolpaw *et al.*, "Brain-computer interface technology: a review of the first international meeting," *IEEE Trans. Rehabil. Eng.*, vol. 8, no. 2, pp. 164–173, 2000.

[4] G. Liberati *et al.*, "Toward a brain-computer interface for Alzheimer's disease patients by combining classical conditioning and brain state classification," *J. Alzheimers Dis.*, vol. 31, no. s3, pp. S211–S220, 2012.

[5] F. Lieder and T. L. Griffiths, "Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources," *Behav. Brain Sci.*, vol. 43, 2020.

[6] L. R. Hochberg *et al.*, "Neuronal ensemble control of prosthetic devices by a human with tetraplegia," *Nature*, vol. 442, no. 7099, pp. 164–171, 2006.

[7] J. K. Chapin, K. A. Moxon, R. S. Markowitz, and M. A. Nicolelis, "Real-time control of a robot arm using simultaneously recorded neurons in the motor cortex," *Nat. Neurosci.*, vol. 2, no. 7, pp. 664–670, 1999.

[8] F. Wörgötter, W. J. Daunicht, and R. Eckmiller, "An on-line spike form discriminator for extracellular recordings based on an analog correlation technique," *J. Neurosci. Methods*, vol. 17, no. 2–3, pp. 141–151, 1986.

[9] H. Kaneko, S. S. Suzuki, J. Okada, and M. Akamatsu, "Multineuronal spike classification based on multisite electrode recording, whole-waveform analysis, and hierarchical clustering," *IEEE Trans. Biomed. Eng.*, vol. 46, no. 3, pp. 280–290, 1999.

[10] S. Mukhopadhyay and G. C. Ray, "A new interpretation of nonlinear energy operator and its efficacy in spike detection," *IEEE Trans. Biomed. Eng.*, vol. 45, no. 2, pp. 180–187, 1998.

[11] K. H. Kim and S. J. Kim, "Neural spike sorting under nearly 0-dB signal-to-noise ratio using nonlinear energy operator and artificial neural-network classifier," *IEEE Trans. Biomed. Eng.*, vol. 47, no. 10, pp. 1406–1411, 2000.

[12] E. Hulata, R. Segev, and E. Ben-Jacob, "A method for spike sorting and detection based on wavelet packets and Shannon's mutual information," *J. Neurosci. Methods*, vol. 117, no. 1, pp. 1–12, 2002.

[13] K. G. Oweiss and D. J. Anderson, "Noise reduction in multichannel neural recordings using a new array wavelet denoising algorithm," *Neurocomputing*, vol. 38, pp. 1687–1693, 2001.

[14] Z. Zhao, X. Li, F. He, X. Wei, S. Lin, and C. Xie, "Parallel, minimally-invasive implantation of ultra-flexible neural electrode arrays," *J. Neural Eng.*, vol. 16, no. 3, p. 035001, 2019.

[15] A. Maccione, M. Gandolfo, P. Massobrio, A. Novellino, S. Martinoia, and M. Chiappalone, "A novel algorithm for precise identification of spikes in extracellularly recorded neuronal signals," *J. Neurosci. Methods*, vol. 177, no. 1, pp. 241–249, 2009.

[16] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[18] D. Yu and L. Deng, *Automatic speech recognition*, vol. 1. Springer, 2016.

[19] M. Han *et al.*, "Improving End-to-End Contextual Speech Recognition with Fine-Grained Contextual Knowledge Selection," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 8532–8536.

[20] K. Chowdhary, "Natural language processing," *Fundam. Artif. Intell.*, pp. 603–649, 2020.

[21] D. Zhang, X. Chen, S. Xu, and B. Xu, "Knowledge aware emotion recognition in textual conversations via multi-task incremental transformer," in *Proceedings of the 28th International Conference on Computational Linguistics*, 2020, pp. 4429–4440.

[22] S. Jia, T. Zhang, X. Cheng, H. Liu, and B. Xu, "Neuronal-plasticity and reward-propagation improved recurrent spiking neural networks," *Front. Neurosci.*, vol. 15, p. 654786, 2021.

[23] T. Zhang, X. Cheng, S. Jia, M. Poo, Y. Zeng, and B. Xu, "Self-backpropagation of synaptic modifications elevates the efficiency of spiking and artificial neural networks," *Sci. Adv.*, vol. 7, no. 43, p. eabh0146, 2021.

[24] A. G. Schwing and R. Urtasun, "Fully connected deep structured networks," *ArXiv Prepr. ArXiv150302351*, 2015.

[25] J. Yang *et al.*, "Quantization networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7308–7316.