NMBNN: Noise-Adaptive Memristive Bayesian Neural Network for Energy-Efficient Edge Health Care

Hanrui Li[1,2], Fengshi Tian[2], Jie Yang[2], Mohamad Sawan[2], Nazek El-Atab[1]
[1]*SAMA Labs, CEMSE department, KAUST, Thuwal, Saudi Arabia, 23955-6900*
[2]*CenBRAIN Neurotech, School of Engineering, Westlake University, Hangzhou, China, 310030*
nazek.elatab@kaust.edu.sa, sawan@westlake.edu.cn

***Abstract*—Energy-efficient and noisy-adaptive signal processing system are in high demand of edge biomedical applications. In this paper, we present a Noise-Adaptive Memristive Bayesian Neural Network (NMBNN) architecture for various biosignal applications. The memristor has the inherent physical property of exhibiting variability in resistance, which makes it a promising candidate of uncertainty weight in Bayesian Neural Networks (BNN). The NMBNN architecture combines the noise-resilient attributes of BNN with the implementation of an energy-efficient RRAM array. By utilizing BNN's probabilistic predictions and implementation with the conductance fluctuations of memristors, NMBNN offers a robust and energy-efficient solution adept at processing biosignals in noisy environments. In order to evaluate the network robustness, we conduct the experiments to introduce multiple types of noise as adversarial sample. The experimental results indicate that the proposed NMBNN approach has the advantages of being both noise-adaptive and energy-efficient.**

***Index Terms*—bayesian neural network, network robustness, memristor, signal processing**

I. Introduction

Artificial intelligence (AI) has greatly contributed to health monitoring and disease diagnosis by utilizing biomedical signals, leading to the development of various edge biomedical applications [1]–[3]. However, many challenges still remain in their practical application. In daily monitoring, acquiring biomedical signal typically requires specialized medical equipment and trained staff, making it difficult to implement in real-world scenarios. Additionally, the signal acquisition equipment inevitably introduces unwanted noise, which can negatively impact network performance and affect the accuracy of prediction models [4]. These limitations motivate the development of new approaches that can deliver robust performance while being energy-efficient and practical for real-world implementation.

Unlike traditional neural networks with point estimates for outputs, the Bayesian neural network (BNN) generates probabilistic predictions, offering the inherent advantage of effectively handling weight uncertainty and dealing with external noise [5]. This robustness makes BNN a promising approach for processing biomedical signals that are subject to various forms of noise and variability. Recently, the progress of resistive random-access memory (RRAM), which is a two-terminal device that contains resistive switching property, provides the potential to use as energy-efficient deep learning accelerators at the edge [6]–[8]. Due to the stochastic ion behavior and vacancy forming process, the conductance of memristors displays fluctuations and uncertainty perturbations [9]. The varying conductive filaments provide variations from the expected weight value and measurement errors, which is similar to the weight uncertainty property in BNN [10]. Therefore, RRAM-based BNN could be a promising candidate for efficient and robust deep-learning applications.

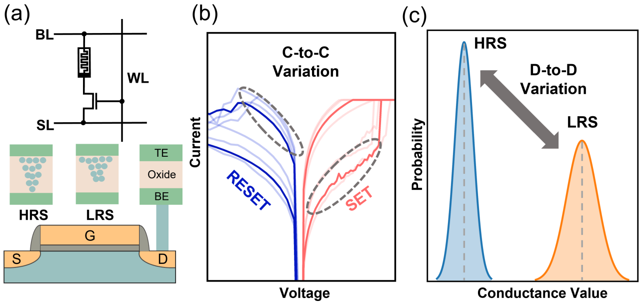

Fig. 1. (a) 1T1M architecture. (b) RRAM-based set and reset curve with cycle-to-cycle variations. (c) During programming, HRS and LRS suffer deviation from the excepted weight with device-to-device variations.

In this paper, we propose the NMBNN architecture, which combines the noise-adaptive capabilities of BNN with the energy efficiency of memristor-based hardware, to overcome the limitations of computation resources and noisy perturbation in real-time assistance. The unique weight uncertainty property of BNN enables it to handle noisy perturbations effectively and perform robust computation in real-world scenarios. We introduce multiple types of noise including Gaussian noise, impulse noise, and signal distortion to test dataset as adversarial samples, and conduct ablation experiments to assess the network's resilience under different noise conditions.

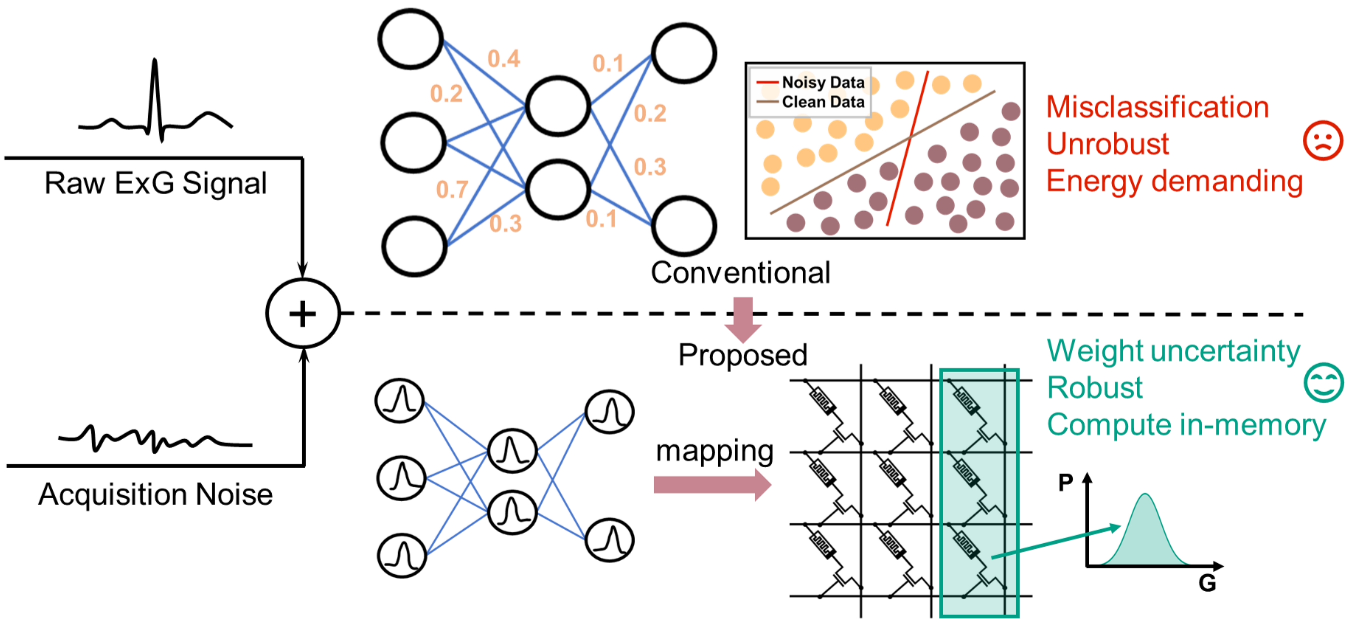

Fig. 2. The theory of proposed method: conventional method is easily affected by noisy perturbation while RRAM-based BNN with the advantage of robustness and weight uncertainty have better performance in the real-world noisy scenario. P: probability; G: conductance value.

II. Preliminaries

A. RRAM-based neural network

As an emerging device, RRAM can efficiently accelerate vector-matrix multiplications in deep learning models by realizing the in-memory computing architecture [9]. The 1T1M configuration is one of the widely employed methods [11], where the transistor serves as a selector to control resistance switching as shown in Fig. 1(a). Through the process of shifting between high resistance state (HRS) to low resistance state (LHS), it would form multiple conductance states, which are well-suited for representing synaptic weights in neural networks. Nonetheless, ion motion randomness or fluctuating conductance filaments invariably result in variations, such as cycle-to-cycle and device-to-device variations, as depicted in Fig. 1(b) and (c). Those variations generate notable deviations from the expected weight value and result in measurement errors, which degrade the performance in conventional neural networks [12].

B. Bayesian neural network

BNNs have gained considerable attention in the machine learning algorithms. Unlike conventional neural network with fixed weights, it offers a probabilistic approach to weight optimization [5]. BNN maintains output as a probability distribution over possible weights, encapsulating an inherent uncertainty associated with the weights of a neural network. The core of a BNN aims to approximate the posterior distributions of weight space $p(w|D)$ with Bayes' theorem:

$$p(w|D) = \frac{p(w)p(D|w)}{p(D)} \tag{1}$$

where $p(D)$ represents the marginal likelihood, $p(w)$ represents the prior weight distribution, and $p(D|w)$ represents the likelihood of observing the data D given the weight w. Due to the intractability of the posterior $P(w|D)$, variational inference proposes the use of a variational distribution $q(w|\theta)$ to approximate the posterior, thus avoiding the intractable computation. This can be accomplished by minimizing the closeness between the two distributions through Kullback-Leibler (KL) divergence:

$$\theta^* = \underset{\theta}{argmin}\; KL[q(w|\theta) \parallel P(w)] - \mathbb{E}_{q(w|\theta)}[\log P\,(D|\theta)] \tag{2}$$

The backpropagation algorithm combines the cost of KL divergence $KL[q(w|\theta) \parallel P(w)]$ and likelihood probability $\mathbb{E}_{q(w|\theta)}[\log P\,(D|\theta)]$. For the priori knowledge $P(w)$ from the above equation, it can be formulated as a Gaussian distribution with the mean μ_i and standard deviation σ_i:

$$P(w) = \mathcal{N}(\mu_i, \sigma_i) \tag{3}$$

For the posterior weight distribution, it aims to find optimal variational parameters $\theta = (\mu, \delta)$ according to the information of priori knowledge. The variation property in RRAM-based network implementation is suitable for exploiting weight uncertainty in BNNs [9], [13].

III. Theory of the proposed approach

A. System architecture

Fig. 2 shows the proposed NMBNN architecture. In daily monitoring, noise is a common issue that can interfere with the quality and reliability of biomedical signals. As raw ExG signals are collected and processed by medical equipment, there is a possibility for the unintended introduction of extraneous noise in acquisition process. For conventional CNN method with fixed weights, it would face challenges in maintaining accuracy when data is subjected to noise-induced perturbations.

In contrast, the proposed NMBNN approach exploits the variability inherent in the memristor crossbar to provide a priori situations to a BNN. It effectively serves as the uncertain weights in the neural network, thereby offering robustness and adaptability to noise. Furthermore, unlike conventional CMOS processors, RRAM have the unique property to store and process the data with computate-in-memory (CIM) architecture, which enhances computational efficiency and reduces the power consumption. By employing the proposed method, the network attains substantially improved performance when handling adversarial noisy samples in the test data. By addressing potential challenges associated with the harmonization of software and hardware requirements, the proposed NMBNN architecture can be further refined for practical implementation.

IV. Experiment evaluation

A. *Noise behaviour analysis*

Noise is a prevalent issue encountered during biomedical signal acquisition and translational shifts procedure due to random fluctuations in physiological processes or environmental factors. One common way to represent noise in biomedical signals is through an additive noise model [14], where the noise component is added to the original clean signal. This can be mathematically formulated as:

$$N(t) = A(t) + n(t) \tag{4}$$

where $N(t)$ represents the noisy signal at time t, $A(t)$ represents the orginal clean signal at time t and $n(t)$ is noisy component.

We consider three common noise phenomenons relevant to biosignal processing: Gaussian noise, impulsive noise, and noisy fold behavior. Gaussian noise, also known as white noise or thermal noise, follows a Gaussian distribution with a zero mean and constant standard deviation. It is caused by random fluctuations in signal and system components, such as thermal noise in electronic devices. Mathematically, Gaussian noise $n_g(t)$, with its mean as μ and standard deviation as σ, can be represented as:

$$n_g(t) = \mu + \sigma\xi(t) \tag{5}$$

where $\xi(t)$ is a random variable following a standard normal distribution.

Impulsive noise is characterized by sudden high-amplitude spikes that appear randomly in the signal and introduces abrupt changes to the analysis process. This noise can be modeled using a Bernoulli-Gaussian sequence [15]:

$$n_i(t) = r(t)q(t) \tag{6}$$

where $r(t)$ is a Bernoulli random variable with probability p of taking the value 1 and probability $(1-p)$ of taking the value 0. $q(t)$ represents the amplitude of the impulse.

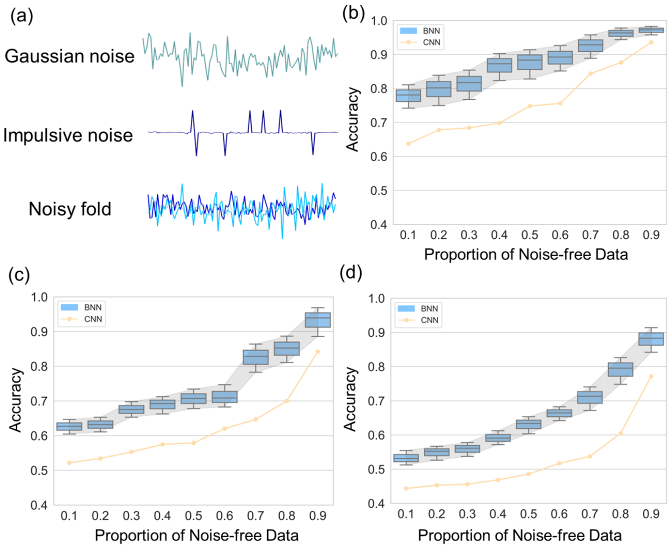

Fig. 3. (a) Demonstration of three kinds of noise. (b) Performance comparison for different proportion of clean data with Gaussian noise. (c) Performance comparison for different proportion of clean data with impulsive noise. (d) Performance comparison for different proportion of clean data with noisy fold behaviour.

For the noise fold situation, it represents a scenario where the noise is complex and folded over itself in multiple layers. The intricate layering makes it tough to discern a clear pattern or feature. We formulate it as a weighted sum of several Gaussian noise components:

$$n_f(t) = \sum_{k=1}^{K} w_k n_{g_k}(t) \tag{7}$$

where K is the total number of Gaussian noise components and w_k represents the weight of the k-th Gaussian noise component.

Fig. 3(a) illustrates these three types of noise phenomena. From this figure, one can observe the distinct characteristics of Gaussian noise, impulsive noise, and multi-component Gaussian noise. After we trained the model, we introduce these types of noise into the test dataset as adversarial samples. The inclusion of adversarial samples with noise perturbation helps assess the model's ability to deal with real-world challenges. This is critical for evaluating the model's robustness and performance under conditions that mimic actual biomedical signal processing scenarios.

B. *Experimental result*

We evaluate our NMBNN model along with noisy analysis for three different biomedical tasks, including EEG-based seizure prediction, EMG-based gesture recognition and ECG-based arrhythmia detection. Our experiment is based on three open-source datasets, including MIT-BIH dataset for ECG, Nina Pro DB1 dataset for EMG and CHB-MIT dataset for EEG. We employ the CNN model directly from paper [16] as a baseline method, and conduct experiments by introducing

TABLE I
PERFORMANCE COMPARISON UNDER MULTIPLE NOISY SITUATIONS

	Dataset	Task	Network type	Acc	Noise type	NP (%)
ECG	MIT-BIH	Arrhythmia detection	5Conv+2FC	98.3	Gaussian	84.38 **92.86**
					Impulsive	64.69 **82.46**
					Noisy fold	51.72 **66.47**
EMG	Nina Pro DB1	Gesture recognition	4Conv+FC	83.14	Gaussian	65.64 **73.64**
					Impulsive	61.42 **68.78**
					Noisy fold	57.24 **67.23**
EEG	CHB-MIT	Epileptic seizure detection	6Conv+2FC	94.23	Gaussian	78.11 **88.53**
					Impulsive	72.21 **84.53**
					Noisy fold	74.23 **85.36**

Acc: accuracy; NP: noisy performance; Conv: convolutional layer; FC: fully connected layer.

TABLE II
COMPARISON WITH PRIOR PUBLICATIONS

	This work	TBioCAS'23 [18]	IET'20 [19]	TBioCAS'20 [19]
Network type	BNN	CNN	SRNN	MLP
Technology node	65nm	55nm	55nm	180nm
Device type	RRAM	CMOS	CMOS	CMOS
Energy (μj/classification)	0.142	0.99	1.99	3.21
Area(mm^2)	2.72	5.06	4	0.93
Performance	**98.49%**	99.38%	97.80%	98.00%
Noise performance	**92.86%**	84.37%	86.45%	90.40%

noisy perturbation. Figures 3(b), (c), and (d) present performance comparisons on the MIT-BIH dataset between CNN and BNN with the introduction of Gaussian noise, impulsive noise, and noisy fold behavior, respectively.

We carry out twenty iterations on test dataset with the weight uncertainty BNN to evaluate the model's boundary and compute the average performance. As expected, both networks exhibited a decrease in classification accuracy as the proportion of noisy signals increased. However, despite the overall decrease in accuracy, BNN demonstrates more stability and robustness compared to conventional CNN model. In an environment with Gaussian noise, CNN experiences a substantial drop in accuracy from 93.59% to 74.84% as the noise proportion increases from 0.1 to 0.5. In comparison, the BNN model exhibits a significantly less pronounced decrease, with an average accuracy reduction of 8.9%, from an initial accuracy of 97.27% to 88.37%. It is evident that RRAM-based BNN maintains robust performance and provides a probability range for test results, which highlight the advantages of NMBNN in terms of noise adaptiveness and robustness.

Table I compares the proposed RRAM-based BNN method with other CNN methods. In terms of noisy performance, the average accuracy of BNN under perturbation is highlighted in bold, while the original CNN method is grayed out for comparison. The baseline CNN method without any noise is listed in Acc column. For EEG-based seizure detection task, the baseline CNN model reaches the accuracy of 94.23% and NMBNN method hold the stable performance of 88.53% with Gaussian noise, which is much higher than the original CNN model (78.11%). For each biomedical task, we can observe the significant noisy influence for original method, which also indicates the challenge of real-time signal processing in edge computing. However, with the introduction of our NMBNN method, we can enhance the model's robustness and lessen the impact of different types of noise on performance. The NMBNN's ability to sustain high levels of accuracy under noisy conditions suggests its potential for deployment in real-time health monitoring systems.

C. Performance evaluation

The proposed NMBNN is evaluated according to the NeuroSim platform [17]. In order to compare the performance of proposed architecture, we estimate the network in 128x128 memristor array for edge ECG-based arrhythmia detection. Except the RRAM array, the NMBNN system includes peripheral circuits with CIM architecture using 65nm technology node, which can deploy a 7-layer network.

Table I make a brief comparison with other state-of-the-art work including area, energy consumption, and detection accuracy. The proposed method exhibits great robustness and energy efficiency in edge ECG-based arrhythmia detection task. Specifically, it achieves a high noise adaptive accuracy of 92.86% and exhibits up to 8 times greater energy efficiency compared to other methods. These results underline the robustness and energy efficiency of the proposed NMBNN, marking it as a promising approach for practical biomedical applications.

V. CONCLUSION

In this paper, we present NMBNN architecture, which can be an energy-efficient and noise-adaptive solution for edge AI biomedical application. By synergistically integrating the probabilistic and noise-resilient attributes of BNN with hardware-friendly implementation of RRAM, the NMBNN method outperforms the conventional CNN based method among three biomedical signal datasets. By effectively handling noisy environments and operating with high energy efficiency, NMBNN paves the way for timely interventions and effective signal processing in critical scenarios.

ACKNOWLEDGMENT

Authors acknowledge financial support from the Semiconductor Initiative, King Abdullah University of Science and Technology, Saudi Arabia (KAUST Research Funding (KRF) under Award No. ORA2022-5314).

REFERENCES

[1] J. Pavei, R. G. Heinzen, B. Novakova, R. Walz, A. J. Serra, M. Reuber, A. Ponnusamy, and J. L. Marques, "Early seizure detection based on cardiac autonomic regulation dynamics," *Frontiers in physiology*, vol. 8, p. 765, 2017.

[2] K. Yu, L. Tan, L. Lin, X. Cheng, Z. Yi, and T. Sato, "Deep-learning-empowered breast cancer auxiliary diagnosis for 5gb remote e-health," *IEEE Wireless Communications*, vol. 28, no. 3, pp. 54–61, 2021.

[3] T. Alafif, A. M. Tehame, S. Bajaba, A. Barnawi, and S. Zia, "Machine and deep learning towards covid-19 diagnosis and treatment: survey, challenges, and future directions," *International journal of environmental research and public health*, vol. 18, no. 3, p. 1117, 2021.

[4] S. Chatterjee, R. S. Thakur, R. N. Yadav, L. Gupta, and D. K. Raghuvanshi, "Review of noise removal techniques in ecg signals," *IET Signal Processing*, vol. 14, no. 9, pp. 569–590, 2020.

[5] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra, "Weight uncertainty in neural network," in *International conference on machine learning*, pp. 1613–1622, PMLR, 2015.

[6] W. Zhang, B. Gao, J. Tang, P. Yao, S. Yu, M.-F. Chang, H.-J. Yoo, H. Qian, and H. Wu, "Neuro-inspired computing chips," *Nature electronics*, vol. 3, no. 7, pp. 371–382, 2020.

[7] B. Gao, B. Lin, Y. Pang, F. Xu, Y. Lu, Y.-C. Chiu, Z. Liu, J. Tang, M.-F. Chang, H. Qian, *et al.*, "Concealable physically unclonable function chip with a memristor array," *Science Advances*, vol. 8, no. 24, p. eabn7753, 2022.

[8] D. Kumar, H. Li, U. K. Das, and A. M. SyedEl-Atab, "Flexible solution processable black phosphorus based optoelectronic memristive synapse for neuromorphic computing and artificial visual perception applications," *Advanced Materials*, p. 2300446, 2023.

[9] D. Ielmini and H.-S. P. Wong, "In-memory computing with resistive switching devices," *Nature electronics*, vol. 1, no. 6, pp. 333–343, 2018.

[10] K.-E. Harabi, T. Hirtzlin, C. Turck, E. Vianello, R. Laurent, J. Droulez, P. Bessière, J.-M. Portal, M. Bocquet, and D. Querlioz, "A memristor-based bayesian machine," *Nature Electronics*, vol. 6, no. 1, pp. 52–63, 2023.

[11] Y. Y. Chen, M. Komura, R. Degraeve, B. Govoreanu, L. Goux, A. Fantini, N. Raghavan, S. Clima, L. Zhang, A. Belmonte, *et al.*, "Improvement of data retention in hfo 2/hf 1t1r rram cell under low operating current," in *2013 IEEE International Electron Devices Meeting*, pp. 10–1, IEEE, 2013.

[12] W.-Q. Pan, J. Chen, R. Kuang, Y. Li, Y.-H. He, G.-R. Feng, N. Duan, T.-C. Chang, and X.-S. Miao, "Strategies to improve the accuracy of memristor-based convolutional neural networks," *IEEE Transactions on Electron Devices*, vol. 67, no. 3, pp. 895–901, 2020.

[13] A. Sebastian, R. Pendurthi, A. Kozhakhmetov, N. Trainor, J. A. Robinson, J. M. Redwing, and S. Das, "Two-dimensional materials-based probabilistic synapses and reconfigurable neurons for measuring inference uncertainty using bayesian neural networks," *Nature communications*, vol. 13, no. 1, p. 6139, 2022.

[14] C. J. James and C. W. Hesse, "Independent component analysis for biomedical signals," *Physiological measurement*, vol. 26, no. 1, p. R15, 2004.

[15] J. M. Leski, "Robust weighted averaging [of biomedical signals]," *IEEE Transactions on Biomedical Engineering*, vol. 49, no. 8, pp. 796–804, 2002.

[16] F. Tian, J. Yang, S. Zhao, and M. Sawan, "Neurocare: A generic neuromorphic edge computing framework for healthcare applications," *Frontiers in Neuroscience*, vol. 17, p. 1093865, 2023.

[17] X. Peng, S. Huang, H. Jiang, A. Lu, and S. Yu, "Dnn+ neurosim v2. 0: An end-to-end benchmarking framework for compute-in-memory accelerators for on-chip training," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 40, no. 11, pp. 2306–2319, 2020.

[18] C. Fang, C. Wang, S. Zhao, F. Tian, J. Yang, and M. Sawan, "A 510μw 0.738-mm2 6.2-pj/sop online learning multi-topology snn processor with unified computation engine in 40-nm cmos," *IEEE Transactions on Biomedical Circuits and Systems*, 2023.

[19] Y. Zhao, Z. Shang, and Y. Lian, "A 13.34 μw event-driven patient-specific ann cardiac arrhythmia classifier for wearable ecg sensors," *IEEE transactions on biomedical circuits and systems*, vol. 14, no. 2, pp. 186–197, 2019.