The following publication S. Cai, R. Zhang, M. Zhang, J. Wu and H. Li, "EEG-Based Auditory Attention Detection With Spiking Graph Convolutional Network," in IEEE Transactions on Cognitive and Developmental Systems, vol. 16, no. 5, pp. 1698-1706, Oct. 2024 is available at https://doi.org/10.1109/TCDS.2024.3376433.

EEG-based Auditory Attention Detection with Spiking Graph Convolutional Network

Siqi Cai *Member, IEEE*, Ran Zhang, Malu Zhang *Member, IEEE*, Jibin Wu *Member, IEEE*, and Haizhou Li, *Fellow, IEEE*

***Abstract*—Decoding auditory attention from brain activities, such as electroencephalography (EEG), sheds light on solving the machine cocktail party problem. However, effective representation of EEG signals remains a challenge. One of the reasons is that the current feature extraction techniques have not fully exploited the spatial information along the EEG signals. EEG signals reflect the collective dynamics of brain activities across different regions. The intricate interactions among these channels, rather than individual EEG channels alone, reflect the distinctive features of brain activities. In this study, we propose a spiking graph convolutional network, called SGCN, which captures the spatial features of multi-channel EEG in a biologically plausible manner. Comprehensive experiments were conducted on two publicly available datasets. Results demonstrate that the proposed SGCN achieves competitive auditory attention detection (AAD) performance in low-latency and low-density EEG settings. As it features low power consumption, the SGCN has the potential for practical implementation in intelligent hearing aids and other BCIs.**

***Index Terms*—Auditory attention, EEG, graph convolutional network, spiking neural network**

I. Introduction

Humans can selectively attend to the speaker of interest in the presence of multiple speakers, which is known as "*cocktail party effect*" [1]. However, listeners with hearing impairment (HI) often struggle to follow the interested speech in such noisy conditions. Despite significant advancements in hearing devices, their potential benefits for those with HI remain limited as these devices lack a direct connection with the human brain. The current generation of hearing devices exhibits a deficiency in effectively identifying the target sound sources that require attention. Specifically, the challenge of discerning speech in complex auditory environments, such as cocktail party scenarios, remains a significant concern for individuals utilizing hearing aids [2]. Addressing this challenge necessitates a collaborative effort between artificial intelligence (AI) and the audiology community to propel a transformative advancement in hearing technology.

Recent advances in neuroscience have shown that the neural patterns associated with auditory attention can be discerned through the analysis of brain activity, such as electrocorticography (ECoG) [3], magnetoencephalography (MEG) [4], and electroencephalography (EEG) [5], [6], etc. That is referred to as "auditory attention detection (AAD)". The development of AAD has shed new light on smart hearing devices, which are called "neuro-steered hearing aids" [2]. By putting humans in the loop, the hearing device is expected to extract and enhance the attended speech as decoded from the listener's brain. Among various techniques for measuring brain activity, EEG demonstrates the advantages of low cost, wide availability, and ease of use, making it a feasible choice for integration into practical brain-computer interface (BCI) applications.

Neuroscience findings indicate that spatially separated brain areas are involved in the selective listening task [7], [8]. Spatial patterns of brain responses to auditory stimuli play a key role in detecting auditory attention. Inspired by this, the common spatial pattern (CSP) method has been adopted in EEG-enabled AAD for spatial feature extraction [9]. With the advent of deep learning, recent research has explored the feasibility of identifying auditory attention from EEG using neural networks and achieved promising results [10], [11]. For instance, Vandecappelle et al. [12] used a CNN model to extract spatial features to determine the spatial locus of auditory attention (i.e., the directional focus of attention, e.g., left or right), solely from EEG data. This approach has the advantage of not requiring individual speech envelopes. Additionally, it avoids the need to estimate a correlation coefficient over a relatively long decision window length, thereby mitigating significant algorithmic delays [13]. Building on this, various CNN variations have been developed [14], and a comprehensive overview can be found in [6]. Despite the success of CNNs in learning spatial information from EEG signals for AAD tasks, it's important to note that CNNs are inherently suitable for Euclidean space and may have limitations when dealing with signals exhibiting a complex topological structure [15]. Indeed, The configuration of multi-channel EEG is a typical example of the irregular layout, with

Manuscript received ; revised . This work is supported by A*STAR under its RIE 2020 Advanced Manufacturing and Engineering Human (AME) Programmatic Grant (Grant No. A1687b0033). The research is also supported by the National Natural Science Foundation of China (Grant No. 62271432 and 62106038); Internal Project of Shenzhen Research Institute of Big Data (Grant No. T00120220002); Shenzhen Science and Technology Research Fund (Fundamental Research Key Project Grant No. JCYJ20220818103001002). The work by Haizhou Li is also funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy (University Allowance, EXC 2077, University of Bremen, Germany).

Siqi Cai and Haizhou Li are with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore. Haizhou Li is also with the Shenzhen Research Institute of Big Data, School of Data Science, The Chinese University of Hong Kong, Shenzhen, China, and the Machine Listening Lab, University of Bremen, Germany.

Ran Zhang is with the Shien-Ming Wu School of Intelligent Engineering, South China University of Technology, China.

Malu Zhang is with the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 610054, China

Jibin Wu is with the Department of Computing, The Hong Kong Polytechnic University, Hong Kong SAR

Corresponding author: Malu Zhang

the spatial and functional connectivity among EEG electrodes containing valuable information [16]. To tackle this challenge, we adopt a graph-based description method for multi-channel EEG signals, aiming to capture the intricate spatial relationships across the whole brain.

Moreover, most deep learning AAD models fall into traditional neural architectures, presenting challenges in terms of energy consumption, data requirements, and computational costs [17]. For edge computing platforms, such as neuro-steered hearing devices, there remains a need for low-power consumption solutions capable of on-chip learning. As well known, the human brain is the most efficient computing machine, which consists of tens of billions of neurons and trillions of synapses connecting them [18]. Information is transmitted between neurons by electrical impulses called spikes. Motivated by the remarkable processing power of the human brain, a biorealistic neural architecture, spiking neural networks (SNNs), has been developed [19]. To mimic the human brain, neurons communicate with each other using spikes with adjustable weight values through synapses connecting neurons in a typical SNN [20]. Unlike traditional neural networks, SNNs process information in the form of spikes or action potentials, which are short-lived events that are generated in response to stimuli. SNNs are designed to be event-driven, meaning that they only update and process information when an event (such as a spike) occurs. This property makes it easy to deploy and achieve ultra-low power consumption on neuromorphic chips [21]–[23]. Due to the great progress in effectively training SNNs [24]–[28], SNNs have been successfully applied in areas such as computer vision, speech recognition, and robotics and have achieved competitive results with low latency and low power consumption [29]–[36]. These findings and results motivate us to study an SNN-based AAD architecture for neuro-steered hearing devices.

Overall, we build a spiking graph convolutional network (SGCN) to detect auditory attention from EEG signals. The proposed SGCN can aggregate EEG channel information and extract the spatial features in a biologically plausible manner. This is similar to the way humans voluntarily attend to one of the multiple incoming sounds. Consequently, it takes full advantage of an event-driven manner and distributed connection, contributing to superiority in computational efficiency. This work makes three main contributions, which are summarized as follows.

- We provide an alternative solution to EEG-based AAD, i.e. SGCN. It encodes brain signals into spikes and works in an event-driven manner with low energy consumption.
- To the best of our knowledge, this is the first study of a spiking graph convolutional mechanism that models complex information processing in the brain.
- We conducted comprehensive experiments and results indicate the SGCN model achieves competitive AAD accuracy while holding the potential of lower computing cost and low latency in neuromorphic hardware.

The organization of this paper is as follows: In Section II, we present the structure of the proposed SGCN. Section III provides details on the databases, data preprocessing, model training, and evaluation. In Section IV, we report the experiments and analyze the results. The findings are discussed in Section V and the paper is concluded in Section VI.

II. Methods

In this study, we introduce a novel method for determining the spatial locus of auditory attention based solely on EEG data [12], [37]. As illustrated in Fig. 1, our proposed solution, the SGCN, operates in an end-to-end manner, taking raw EEG signals as input and producing classification outputs to identify auditory attention.

A. Graph Convolutional Network

1) EEG Graph: Suppose there are N channels in the EEG input, we take each channel as a node to generate the graph representation. EEG signal $\mathbf{E}_s$ can be transformed into an undirected graph $G = (V, E)$ within a non-Euclidean space. Specifically, V denotes the set of $|V|$ = N nodes, and $(V_i, V_j) \in E$ represents the set of links connecting these channels. $\mathbf{A} \in \mathbb{R}^{N \times N}$ is an adjacency matrix that can be used to express the intrinsic connections between EEG channels. Specifically, the elements of $\mathbf{A}$ are predetermined based on the spatial relationship of the EEG channels [15], [16], as shown in Fig. 2. The entry of the adjacency matrix $a_{i,j}$ measures the level of connection between the channels i and j.

2) Graph Convolution: The traditional CNN model is enhanced by the integration of spectral theory in the design of a graph convolutional network (GCN) [15]. The spectral theory provides a mathematical framework for analyzing the properties of graphs, commonly used to represent complex non-Euclidean data structures. In particular, the graph convolution operation aggregates the features of a vertex and its neighboring vertices to generate a new representation for the vertex [38]. This representation not only captures crucial information about the local structure of the EEG graph but also preserves the topological information about the brain. Such an enhanced representation of the EEG signal is expected to benefit various EEG analysis tasks, including those related to AAD.

As defined in [38], the graph convolution operation is performed by computing the eigendecomposition of the graph Laplacian in the Fourier domain. The Laplacian matrix of a graph can be expressed as:

$$\mathbf{L} = \mathbf{D} - \mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T \tag{1}$$

where $\mathbf{D}$ is the degree matrix, $\mathbf{U}$ is the matrix of eigenvectors, and $\mathbf{\Lambda}$ is the diagonal matrix of its eigenvalues.

The graph convolution operation can be formulated as the multiplication of a signal $\mathbf{x}$ with a filter $g_\theta = diag(\theta)$, parameterized by $\theta \in \mathbb{R}^N$,

$$\mathbf{x}^{'} = \mathbf{U} g_\theta \mathbf{U}^T \mathbf{x} \tag{2}$$

As shown in Fig. 1 (b), we apply a graph convolutional module for representation learning of the EEG graph G.

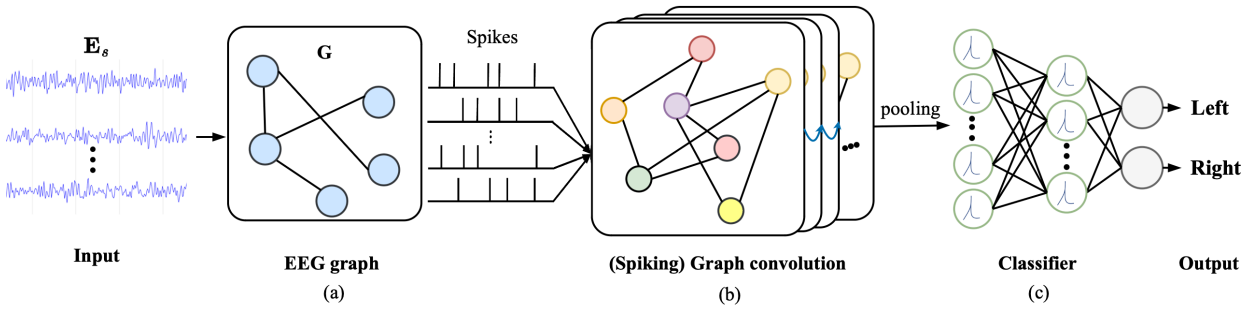

Fig. 1. The framework of our proposed spiking graph convolutional network, i.e., SGCN, for EEG-based AAD. It processes raw EEG signals to detect auditory attention (left/right).

B. Spiking Graph Convolution Network

1) Spiking Neuron Model: Spiking neurons, the essential components of spiking neural networks, exhibit a pivotal role in information processing by transforming inputs into output spike trains tailored for specific tasks. Spike trains are binary time series that encode information in the spiking times. In the input layer, spiking trains are ideally generated by event-driven sensors [39]. Notably, EEG signals exhibit a spiky nature, but they are continuous in value. Therefore, spiking neurons in the input layer are employed to transform the real-valued EEG signals into binary spike trains. This allows for the relevant information in the EEG signals to be represented in a more efficient manner for AAD tasks.

A variety of spiking neuron models have been explored, inspired by the intricate mechanisms of the human nervous system [40], [41]. Among these models, the Leaky Integrate-and-Fire (LIF) neuron model is widely used because it strikes a balance between biological realism and computational efficiency.

The membrane potential V_j^l of LIF neuron j at layer l can be formulated by

$$V_j^l[t] = \lambda V_j^l[t-1] + I_j^l[t] - \vartheta o_j^l[t-1] \tag{3}$$

with

$$I_j^l[t] = \sum_i w_{ji} o_i^{l-1}[t] + b_j^l \tag{4}$$

where λ represents the leak factor, ϑ denotes the firing threshold, $I_j^l[t]$ is the current received from presynaptic neurons by neuron j, w_{ji} is the weight of the connection between presynaptic neuron i and postsynaptic neuron j, and b_j^l denotes the constant current injected into neuron j.

As shown in Fig. 3, the spikes generated by LIF neurons are defined as:

$$o_j^l[t] = \begin{cases} 1, & \text{if } V_j^l[t] > \vartheta, \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

The spike event occurs when the membrane potential reaches the firing threshold ϑ (typically set to $\vartheta = 1$) at time t. After firing, the $V_j[t]$ is reset to the resting potential V_{rest} and remains in a refractory period for a specified duration.

In this study, the real-valued inputs are used as the time-varying input currents and are directly applied in equation 3 at the first time step. The spike count serves as a bridge between the ANN and SNN, as described below:

$$c_i^l = \sum_{t=1}^{N_t} o_i^l[t]. \tag{6}$$

where N_t represents the count of time steps.

2) Spiking Graph Convolution: Spiking graph convolution is a relatively new concept in the field of SNN, as it combines graph convolutional operations with spiking neuronal functions. It leverages the advantages of SNNs, such as event-driven processing and energy efficiency, while incorporating graph structure information into the model. This approach aims to bring the benefits of both SNNs and GCNs to a range of applications, such as computer vision and speech recognition [42].

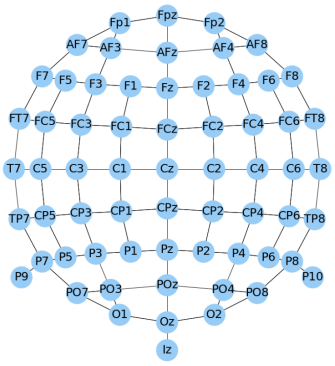

Fig. 2. The topological relationship of 64-channel EEG.

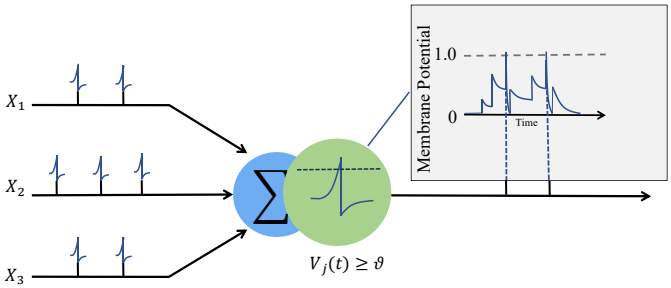

Fig. 3. The Leaky-and-Fire (LIF) spiking neuron model utilized in this study. Both the leaky factor λ and the firing threshold ϑ are all set to 1. The postsynaptic neuron accumulates the spikes from the presynaptic neurons and fires a spike when the membrane potential $V(t)$ reaches the firing threshold.

In practice, the input current will be defined by the graph convolution at each time step:

$$I_j^l[t] = \mathbf{U} g_\theta \mathbf{U}^T \mathbf{x} \tag{7}$$

C. Learning and AAD Classifier

Due to the non-differentiable spike function, how to train efficiently SNNs is a challenging problem. To resolve this problem, various learning algorithms are proposed, such as ANN-to-SNN [26], [27], [43], [44] and surrogate gradient-based methods [24], [45]. Due to the simplicity and efficiency, tandem learning [26], [43] is applied in this work to train the proposed spiking graph convolution neural network.

As depicted in Fig. 1 (c), an SNN classifier is employed for the identification of auditory attention. It comprises three hidden layers, of which two are composed of LIF spiking neurons. The first and second layers consist of 32 and 8 spiking neurons, respectively. The SNN back-end is then applied to decode the output spike counts of the second layer into pattern classes, representing the auditory attention of the listening subjects.

The SGCN model is trained end-to-end and the binary cross-entropy loss serves as the learning objective:

$$\mathcal{L} = -\frac{1}{M} \sum_{m=1}^{M} [y_m \cdot log P_m + (1 - y_m) \cdot log(1 - P_m)] \tag{8}$$

with

$$P_m = softmax(C^{'}) \tag{9}$$

where y_m is the label of the m-th decision window, M is the batch size and $C^{'}$ is the spiking count of the output from the SNN classifier.

III. Experiments

A. Auditory Attention Detection Databases

In this paper, comprehensive experiments are conducted on two publicly available AAD databases, which are summarized in Table I.

In Das-2015 [46], 16 normal-hearing subjects were recruited in the selective listening task. The speech stimuli consist of four Dutch stories, narrated by three male Flemish speakers. The stimuli were either presented dichotically (one speaker per ear) or after head-related transfer function (HRTF) filtering to simulate speech at $\pm$ 90°. Subjects were instructed to selectively attend to one of the two simultaneous speakers. Throughout the experiments, the order of presentation of both conditions was randomized over the different subjects. 64-channel EEG was recorded using a BioSemi ActiveTwo device at a sampling rate of 8,192 Hz. In total, 8x6 min of EEG data was collected for each subject, accumulating 12.8 hours of EEG data for all 16 subjects. This database is referred as to the *KUL* hereafter.

In Fuglsang-2018 [47], 18 subjects with normal hearing selectively attended to one of the two simultaneous speakers. Speech stimuli were excerpts taken from Danish audiobooks and narrated by a male and a female speaker. The stimuli were presented to simulate speech at $\pm$ 60°using HRTF filtering. For each subject, the experiment consisted of 60 trials, each 50 seconds long, for a total of 50 minutes. EEG signals were recorded using a 64-channel BioSemi cap at a sampling frequency of 512 Hz. This database is referred as to the *DTU* hereafter.

B. Data Preprocessing

First, EEG data are re-referenced to the average response of all channels. Second, EEG data are bandpass filtered in the β-band (12-30 Hz) by a 6th-order Chebyshev Type II bandpass filter, and downsampled to 128 Hz. The frequency range is chosen based on the previous AAD studies [12], [37] that β-band is the most informative EEG frequency band to decode the auditory spatial attention. Finally, the EEG data channels were normalized to ensure zero mean and unit variance for each trial. Considering that the proposed SGCN is an end-to-end AAD architecture, no artifact removal operations are involved in the data preprocessing.

For every trial, a sliding window technique is applied to the EEG data to slice the EEG series into smaller durations, hereafter referred to as *decision window*, with an overlap rate of 50%. For each decision window, the EEG data is represented as a graph that reflects the brain's topological structure. Humans are capable of shifting attention from one speaker to another within 1 second [4], and real-world applications require low-latency AAD solutions. This motivates us to focus on shorter decision windows and analyze our AAD model with decision windows of 0.1-second, 0.2-second, 0.5-second, and 1-second, respectively.

C. Training and Evaluation

The AAD models are evaluated using a 5-fold cross-validation (CV) method [48] in a subject-dependent manner. AAD accuracy is calculated as the percentage of correctly detected windows out of all decision windows. The final result reported in this study is the average accuracy obtained from the 5-fold validation process.

As mentioned above, the EEG input is transformed into a graph representation and then processed through a spiking graph convolutional module with trainable weights. Here we present the proposed model with a case study with a 1-second decision window. The EEG data $\mathbf{E}_s \in \mathbb{R}^{128\times64}$, i.e. 128 samples by 64 channels, is encoded into an EEG graph G. In the spiking graph convolutional module, the trainable weight matrix is $g_\theta \in \mathbb{R}^{10\times64\times64}$. Subsequently, a global average pooling layer is used to reduce node features. The data is flattened into a one-dimensional vector, serving as inputs to *fc* layers (input: 640, hidden: 32 and 8, output: 2) designed for auditory attention detection.

A grid search on the validation set is conducted to select appropriate values for all hyperparameters. The network is trained using Adam optimization with a learning rate of 10^{-3}. To improve generalization and prevent overfitting, techniques such as dropout and batch normalization are incorporated.

TABLE I
THE CHARACTERISTICS OF TWO PUBLICLY AVAILABLE AAD DATABASES.

Dataset	# Subjects	Language	Spatial distribution of sound sources	# EEG channels	Duration per subject (min)
Das-2015 [46]	16	Dutch	90° to the left and 90° to the right	64	48
Fuglsang-2018 [47]	18	Danish	60° to the left and 60° to the right	64	50

IV. RESULTS

We first investigate the superiority of graph representation on EEG signals by comparing GCN and CNN on KUL and DTU databases. We then evaluate the AAD accuracy of Spiking GCN, i.e., SGCN, across different decision windows. Finally, we study the impact of EEG channels on the SGCN-based AAD.

A. Analysis of Graph Representation

For a fair comparison, we re-implement the CNN-based AAD model in [12] in our experimental setting. In brief, the CNN model includes a convolution layer with a 64×17 kernel, an average pooling, and two *fc* layers (input: 5, hidden: 5, output: 2), with a sigmoid activation function and a cross-entropy as the loss function. This CNN consists of approximately 5500 parameters. For a fair comparison, we tuned the hyperparameters of the CNN model in the same way we did for our GCN model. For statistical analyses, the normality of the data distribution was assessed using the Kolmogorov-Smirnov test before selecting appropriate statistical tests. To compare the performance differences in AAD between these two models, we utilized paired *t*-tests with a significance level set at 0.05.

As shown in Fig. 4 (a), the CNN model attains an average AAD accuracy of 63.3%, with a standard deviation (SD) of 5.79% with a 1-second decision window on the DTU database. The AAD accuracy of the GCN model is significantly better than that of the CNN model, with a large margin of 9.8% (mean: 73.1%, SD: 7.39%). Similarly, GCN consistently outperforms the CNN model (mean: 80.7%, SD: 9.49%) with an average AAD accuracy of 85.6% (SD: 6.83%) on the KUL database, as shown in Fig. 4 (b). It is worth noting that the AAD performance varies across the subjects. As highlighted in previous studies [6], [49], the spatial origin, amplitude change, and overall variability of brain signals exhibit subject-specific patterns. In addition, factors such as participants' concentration abilities, familiarity with the task, and potential confounding variables contribute to the variability in AAD performance [5].

Significant statistical differences were observed between GCN and CNN (paired *t*-test: p <0.001) on both KUL and DTU datasets. These results support our hypothesis that GCNs can learn more discriminative spatial features of EEG signals than CNN models, thereby improving AAD results. One explanation is that CNNs cannot capture the complex neighborhood information of EEG because they focus on local regions with fixed connections. In contrast, GCNs preserve the brain's rich topological information, leading to a more efficient representation of EEG signals.

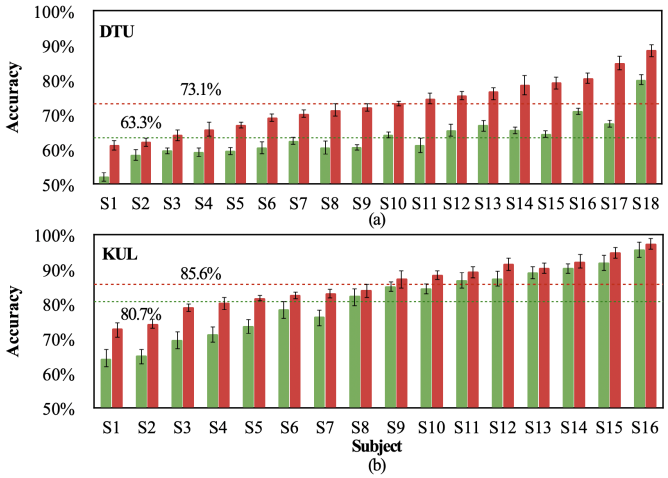

Fig. 4. AAD accuracy (%) of the CNN and GCN model for 1-second decision window with 64-channel EEG on the DTU and KUL databases, respectively.

B. AAD performance of SGCN

As shown in Fig. 5, the SGCN model attains an average AAD accuracy of 68.7% (SD: 6.12%) and 81.4% (SD: 5.95%) with a 1-second decision window on the DTU and KUL database, respectively. Subjects in the DTU database have significantly lower AAD accuracy compared to those in the KUL database, which aligns with previous research findings [9], [12], [50]. One possible explanation could be that the speech streams in the DTU database are delivered at ± 60 °azimuth, whereas the streams come from an azimuthal angle of ± 90 °in the KUL database. This difference may pose a greater challenge in determining the spatial location of the target speaker in the DTU database.

The proposed SGCN performs better than CNN on both KUL and DTU datasets, and yet is inferior to GCN in terms of AAD accuracy. However, the SGCN model may not perform as well as the dense GCN model, but it shines in terms of computational efficiency. This is due to its event-driven approach and distributed connection, which speed up information processing. Despite a slight dip in performance, the SGCN model's exceptional computational efficiency makes it a prime choice for various applications. An in-depth examination of the computational cost will be conducted and discussed in Section IV-E.

C. Effect of EEG Channels

Low-density EEG systems are compact and portable, making them well-suited for real-world applications. To assess the performance of our proposed SGCN model in scenarios with limited data, we conducted additional tests using low-density EEG data.

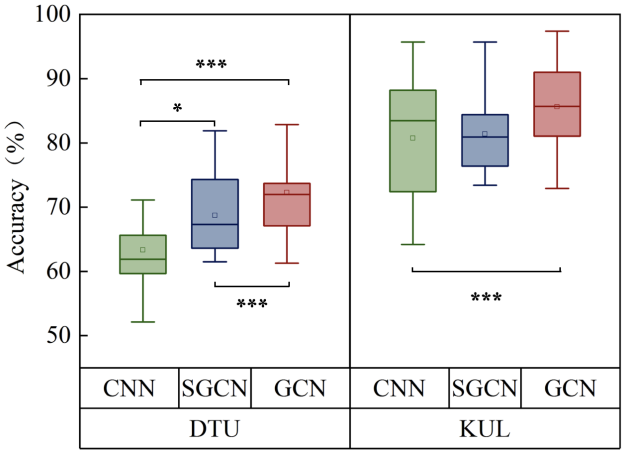

Fig. 5. AAD accuracy (%) of the CNN and GCN model for 1-second decision window with 64-channel EEG on the DTU and KUL databases. Statistically significant differences (paired *t*-tests): $^{*}p$ <0.05, $^{***}p$ <0.001.

Specifically, AAD performance of the proposed SGCN was evaluated using EEG signals from 64-channel, 32-channel, and 16-channel configurations. The KUL dataset provided EEG signals recorded with a 64-channel BioSemi cap, following the international 10/20 system [51]. Additionally, alternative options with 32-channel and 16-channel BioSemi caps, widely used in BCIs, are detailed on the BioSemi website (https://www.biosemi.com/headcap.htm). The results, depicted in Fig. 6, show a general improvement in SGCN performance with an increased number of channels. While more channels provide comprehensive information about brain activity, the selection of an EEG recording system for practical BCI applications should consider the trade-off between channel density, performance, cost, and portability.

It is noteworthy that the AAD performance of the SGCN model under low-density EEG testing is promising, with an average accuracy of 73.1% (SD: 5.19%) attained for 16-channel EEG and a 1-second decision window. This result demonstrates that our proposed SGCN model exhibits the ability to effectively learn and utilize information from fewer EEG electrodes, underscoring its promising potential for practical applications.

D. Effect of Decision Window Sizes

In this study, we also investigated how well the AAD models perform in low-latency settings on the KUL database.

As summarized in Table. II, it can be observed that as the decision window increases, the AAD accuracy generally improves for all three models. Specifically, the SGCN model achieves an average AAD accuracy of 74.7% (SD: 6.04%) for a 0.2-second decision window and 78.0% (SD: 5.41%) for a 0.5-second decision window. Even with a 0.1-second decision window, the SGCN model maintains an acceptable AAD performance with an average accuracy of 70.2% (SD: 5.15%). The accuracy of the SGCN model ranges from 65.7% for a 0.1-second window to 80.7% for a 1-second window. Similarly, the CNN model also exhibits increasing accuracy with longer decision windows. Notably, the results indicate that the SGCN and GCN models may be less sensitive to changes in decision window sizes compared to the CNN model.

In sum, the SGCN and GCN models exhibit promising potential for real-time processing of EEG signals that can quickly respond to changes in inputs.

E. Comparison of Computational Cost

In this section, we further compare the proposed SGCN with traditional neural networks, namely, CNN and GCN models, in terms of computational cost. The total computational cost of a model is proportional to the total number of floating point operations per second (FLOPs) required to make predictions [52]. The standard 45nm CMOS process is a measure of the efficiency of a microprocessor, that is, how many transistors can be packed into a given area. The smaller the process node, the more transistors can be packed into a given area, resulting in higher performance and lower power consumption. When evaluating the total inference cost of a model, the standard 45nm CMOS process can be used as a reference to compare the relative efficiency of different models. This is useful when choosing a model for deployment, as it provides a measure of the computational resources required to run the model, taking into account the efficiency of the hardware used.

It is worth noting that the SGCN model features low power consumption through activating neurons only when sufficient input spikes pass the threshold. Inactive neurons can then be placed in a low-power mode. The model's computation is event-driven by binary spike $\{1, 0\}$ processing, leading to a reduction in computation to just FP addition. In comparison, both CNN and GCN models have low computational efficiency due to the requirement of FP addition and FP multiplication for each MAC operation. As summarized in Table III, the computational cost of our spiking GCN implementation is significantly lower than the GCN implementation (paired *t*-test: p <0.001) in two databases. Compared to GCN, the SNN model achieves an average computational cost reduction of 76.34% and 76.58% in KUL and DTU databases, respectively.

In a nutshell, previous EEG-based AAD architectures are computationally expensive, making them unsuitable for resource-constrained devices, such as mobile or wearable devices. Our proposed SGCN is a promising solution to this problem, as it offers tremendous energy benefits and requires significantly less computing resources and processing power. This makes the SGCN-based AAD architecture more suitable for neuro-steered hearing aids requiring low-power and efficient learning algorithms.

TABLE II
COMPARISON OF AAD ACCURACY (%) ACROSS VARIOUS MODELS FOR DIFFERENT DECISION WINDOW LENGTHS.

AAD Model	Decision window (second)			
	0.1	0.2	0.5	1
CNN [12]	65.7%	70.1%	73.3%	80.7%
GCN	75.2%	80.4%	83.3%	85.6%
SGCN(Ours)	70.2%	74.7%	78.0%	81.4%

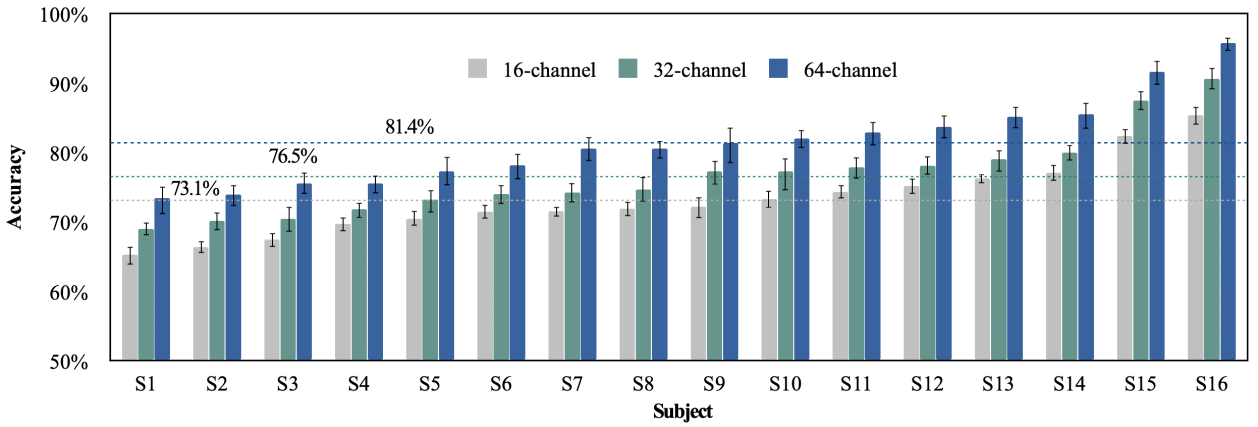

Fig. 6. AAD accuracy (%) of the SGCN model for 1-second decision window with 16-channel, 32-channel, and 64-channel EEG signals on the KUL database.

V. Discussion

Considering that the AAD model is built to process brain signals, a brain-like model should be a natural choice. Results have shown that auditory attention can be well detected by the SGCN model from low-density EEG signals in low-latency settings. We further visualize the spiking representation generated by SGCN to gain more understanding of this biologically inspired AAD decoder.

A. Bio-plausible Visualization of Spiking Learning

As mentioned above, the neuronal mechanism behind how the human brain focuses on interesting speech in a cocktail party environment is also not yet fully understood. Unfortunately, the process by which an ANN reaches a decision is difficult to understand, making it challenging to explain the reasoning behind the output [53]. This lack of interpretability is commonly referred to as the "black box" problem in previous AAD models. To make the AAD model more transparent and understandable, we apply the Spike Activation Map (SAM) technique to enhance understanding and obtain a visual interpretation of the SGCN [54]. SAM computes a neuronal contribution score and generates a 2D spatial heatmap by considering short Inter-Spike-Interval (ISI) spikes as more informative [55]. This highlights neurons that carry significant information for detecting auditory attention across different time steps.

As shown in Fig. 7, several spiking representations of EEG signals generated by SGCN are randomly chosen from different examples. From E1-E4, we can see that the spiking representations gradually converge and show different patterns in leftward and rightward auditory attention conditions. However, in cases where the SGCN's output doesn't match the true label, the spiking representations appear even (E5) or dispersed (E6), which explains the model's incorrect classification of EEG examples. The bio-plausible visualization of EEG representations generated by the SGCN model sheds new light on the neuroscience mechanism underlying auditory attention in the human brain.

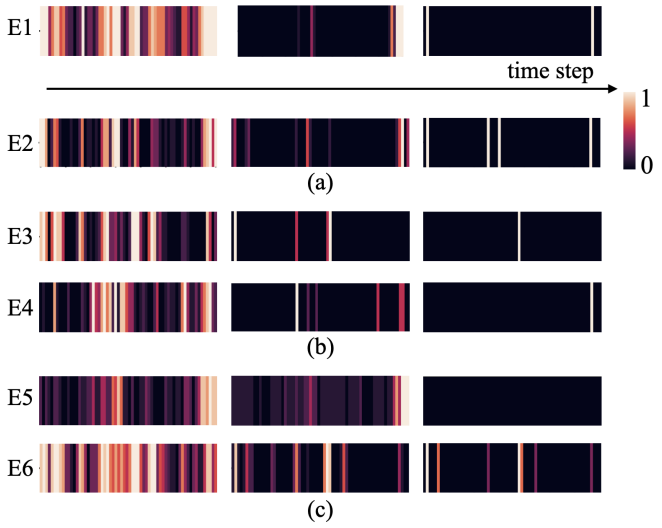

Fig. 7. Visualization of the spiking representations of EEG signals generated by SGCN. Six examples (E1–E6) were randomly selected. E1-E4 are examples of correct classification, whereas E5 and E6 are examples of wrong classification. (a) In E1 and E2, the attended speech stream is on the left of the listening subject. The label and the output of E1 and E2 are "leftward attention". (b) In E3 and E4, the attended speech stream is on the right of the listening subject. The label and the output of E3 and E4 are "rightward attention". (c) E5 and E6 exhibit inconsistent labeling and output. Cell color indicates firing rates, with lighter shades corresponding to more firing rates.

TABLE III
Computational cost (pJ) comparison between SGCN and GCN models on KUL and DTU databases

Database	Computational cost (pJ)		
	E_{GCN}	E_{SGCN}	E_{SGCN} / E_{GCN}
KUL	2.41+E8	5.71+E7	0.2366
DTU	2.41+E8	5.65+E7	0.2342

B. Future work

Our future works will focus on the following aspects. Firstly, we will employ more biologically plausible spiking neuron models to fully take advantage of the spatiotemporal dynamics of biological neurons, such as the Hodgkin-Huxley (HH) model [56]. The introduction of HH spiking neurons will further improve the ability to extract spatiotemporal features

of EEG signals. In addition, we will explore the performance of other graph neural networks within the spiking framework, such as graph attention networks (GAT). Specifically, the capability of GAT to compute learnable attention coefficients for individual EEG channels holds promise for improving accuracy in AAD tasks. Furthermore, we will evaluate the AAD performance in subject-independent conditions, aiming to assess the adaptability and robustness of our proposed approach across diverse individuals. This will involve the integration of advanced transfer learning techniques and the crafting of adaptive models capable of dynamically adjusting to individual differences. Moreover, future research could evaluate the performance of our proposed models on ear EEG, showcasing potential applications in the development of brain-computer interfaces for auditory processing [6]. Unlike conventional scalp EEG methods, utilizing ear EEG offers the advantage of increased convenience and portability, albeit with the trade-off of reduced brain coverage.

VI. Conclusion

In this paper, we present the SGCN, a novel AAD architecture that integrates graph representation and spiking learning. Through validation on two public databases, our model demonstrates competitive AAD performance. Moreover, it leverages a neural computation and coding strategy inspired by the human brain, leading to hardware and energy advantages. Despite the unoptimized performance-complexity trade-off, the proposed SGCN has clear benefits in biological relevance and low power consumption, making it an appealing candidate for future studies to enhance its performance. Moreover, the interpretability of our model contributes to identifying specific areas that necessitate refinement and optimization in future research.

References

[1] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *The Journal of the Acoustical Society of America*, vol. 25, no. 5, pp. 975–979, 1953.

[2] E. Ceolini, J. Hjortkjær, D. D. Wong, J. O'Sullivan, V. S. Raghavan, J. Herrero, A. D. Mehta, S.-C. Liu, and N. Mesgarani, "Brain-informed speech separation (BISS) for enhancement of target speaker in multitalker speech perception," *NeuroImage*, vol. 223, p. 117282, 2020.

[3] N. Mesgarani and E. F. Chang, "Selective cortical representation of attended speaker in multi-talker speech perception," *Nature*, vol. 485, no. 7397, p. 233, 2012.

[4] S. Akram, A. Presacco, J. Z. Simon, S. A. Shamma, and B. Babadi, "Robust decoding of selective auditory attention from MEG in a competing-speaker environment via state-space modeling," *NeuroImage*, vol. 124, pp. 906–917, 2016.

[5] J. A. O'Sullivan, A. J. Power, N. Mesgarani, S. Rajaram, J. J. Foxe, B. G. Shinn-Cunningham, M. Slaney, S. A. Shamma, and E. C. Lalor, "Attentional selection in a cocktail party environment can be decoded from single-trial EEG," *Cerebral Cortex*, vol. 25, no. 7, pp. 1697–1706, 2015.

[6] S. Cai, H. Zhu, T. Schultz, and H. Li, "EEG-based auditory attention detection in cocktail party environment," *APSIPA Transactions on Signal and Information Processing*, vol. 12, no. 3, 2023.

[7] Y. Deng, I. Choi, and B. Shinn-Cunningham, "Topographic specificity of alpha power during auditory spatial attention," *NeuroImage*, vol. 207, p. 116360, 2020.

[8] S. Tune, M. Alavash, L. Fiedler, and J. Obleser, "Neural attentional-filter mechanisms of listening success in middle-aged and older individuals," *Nature Communications*, vol. 12, no. 1, pp. 1–14, 2021.

[9] S. Cai, E. Su, Y. Song, L. Xie, and H. Li, "Low latency auditory attention detection with common spatial pattern analysis of EEG signals," *Proc. Interspeech 2020*, pp. 2772–2776, 2020.

[10] G. Ciccarelli, M. Nolan, J. Perricone, P. T. Calamia, S. Haro, J. O'Sullivan, N. Mesgarani, T. F. Quatieri, and C. J. Smalt, "Comparison of two-talker attention decoding from eeg with nonlinear neural networks and linear methods," *Scientific reports*, vol. 9, no. 1, pp. 1–10, 2019.

[11] T. de Taillez, B. Kollmeier, and B. T. Meyer, "Machine learning for decoding listeners' attention from electroencephalography evoked by continuous speech," *European Journal of Neuroscience*, vol. 51, no. 5, pp. 1234–1241, 2020.

[12] S. Vandecappelle, L. Deckers, N. Das, A. H. Ansari, A. Bertrand, and T. Francart, "EEG-based detection of the locus of auditory attention with convolutional neural networks," *Elife*, vol. 10, p. e56481, 2021.

[13] S.Geirnaert, S. Vandecappelle, E. Alickovic, A. de Cheveigne, E. Lalor, B. T. Meyer, S. Miran, T. Francart, and A. Bertrand, "Electroencephalography-based auditory attention decoding: Toward neurosteered hearing devices," *IEEE Signal Processing Magazine*, vol. 38, no. 4, pp. 89–102, 2021.

[14] E. Su, S. Cai, L. Xie, H. Li, and T. Schultz, "STAnet: A spatiotemporal attention network for decoding auditory spatial attention from EEG," *IEEE Transactions on Biomedical Engineering*, vol. 69, no. 7, pp. 2233–2242, 2022.

[15] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and S. Y. Philip, "A comprehensive survey on graph neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 1, pp. 4–24, 2020.

[16] D. D. Chakladar, P. P. Roy, and M. Iwamura, "EEG-based cognitive state classification and analysis of brain dynamics using deep ensemble model and graphical brain network," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 14, no. 4, pp. 1507–1519, 2021.

[17] A. Taherkhani, A. Belatreche, Y. Li, G. Cosma, L. P. Maguire, and T. M. McGinnity, "A review of learning in biologically plausible spiking neural networks," *Neural Networks*, vol. 122, pp. 253–272, 2020.

[18] F. A. Azevedo, L. R. Carvalho, L. T. Grinberg, J. M. Farfel, R. E. Ferretti, R. E. Leite, W. J. Filho, R. Lent, and S. Herculano-Houzel, "Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain," *Journal of Comparative Neurology*, vol. 513, no. 5, pp. 532–541, 2009.

[19] S. Ghosh-Dastidar and H. Adeli, "Spiking neural networks," *International Journal of Neural Systems*, vol. 19, no. 04, pp. 295–308, 2009.

[20] C. Lee, G. Srinivasan, P. Panda, and K. Roy, "Deep spiking convolutional neural network trained with unsupervised spike-timing-dependent plasticity," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 11, no. 3, pp. 384–394, 2018.

[21] J. Pei, L. Deng, S. Song, M. Zhao, Y. Zhang, S. Wu, G. Wang, Z. Zou, Z. Wu, W. He *et al.*, "Towards artificial general intelligence with hybrid tianjic chip architecture," *Nature*, vol. 572, no. 7767, pp. 106–111, 2019.

[22] M. Davies, N. Srinivasa, T.-H. Lin, G. Chinya, Y. Cao, S. H. Choday, G. Dimou, P. Joshi, N. Imam, S. Jain *et al.*, "Loihi: A neuromorphic manycore processor with on-chip learning," *Ieee Micro*, vol. 38, no. 1, pp. 82–99, 2018.

[23] S. B. Furber, D. R. Lester, L. A. Plana, J. D. Garside, E. Painkras, S. Temple, and A. D. Brown, "Overview of the spinnaker system architecture," *IEEE transactions on computers*, vol. 62, no. 12, pp. 2454–2467, 2012.

[24] E. O. Neftci, H. Mostafa, and F. Zenke, "Surrogate gradient learning in spiking neural networks: Bringing the power of gradient-based optimization to spiking neural networks," *IEEE Signal Processing Magazine*, vol. 36, no. 6, pp. 51–63, 2019.

[25] M. Zhang, J. Wang, J. Wu, A. Belatreche, B. Amornpaisannon, Z. Zhang, V. P. K. Miriyala, H. Qu, Y. Chua, T. E. Carlson *et al.*, "Rectified linear postsynaptic potential function for backpropagation in deep spiking neural networks," *IEEE transactions on neural networks and learning systems*, vol. 33, no. 5, pp. 1947–1958, 2021.

[26] J. Wu, Y. Chua, M. Zhang, G. Li, H. Li, and K. C. Tan, "A tandem learning rule for effective training and rapid inference of deep spiking neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[27] Z. Wang, Y. Zhang, S. Lian, X. Cui, R. Yan, and H. Tang, "Toward high-accuracy and low-latency spiking neural networks with two-stage optimization," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[28] Q. Yang, M. Zhang, J. Wu, K. C. Tan, and H. Li, "Lc-ttfs: Towards lossless network conversion for spiking neural networks with ttfs coding," *IEEE Transactions on Cognitive and Developmental Systems*, 2023.

[29] R. Xiao, Y. Wan, B. Yang, H. Zhang, H. Tang, D. F. Wong, and B. Chen, "Towards energy-preserving natural language understanding with spiking neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 439–447, 2022.

[30] R. Xiao, Q. Yu, R. Yan, and H. Tang, "Fast and accurate classification with a multi-spike learning algorithm for spiking neurons." in *IJCAI*, 2019, pp. 1445–1451.

[31] H. Tang, P. Gu, J. Wijekoon, M. Alsakkal, Z. Wang, J. Shen, and R. Yan, "Neuromorphic auditory perception by neural spiketrum," *arXiv preprint arXiv:2309.05430*, 2023.

[32] L. Qin, R. Yan, and H. Tang, "A low latency adaptive coding spike framework for deep reinforcement learning," in *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, 2023, pp. 3049–3057.

[33] G. Ma, R. Yan, and H. Tang, "Exploiting noise as a resource for computation and learning in spiking neural networks," *arXiv preprint arXiv:2305.16044*, 2023.

[34] K. Liu, J. Shang, X. Cui, C. Zou, Y. Kuang, K. Xiao, Y. Zhong, and Y. Wang, "How the brain achieves real-time vision: A spiking position perception model," *IEEE Transactions on Cognitive and Developmental Systems*, 2023.

[35] L. Guo, Z. Gao, J. Qu, S. Zheng, R. Jiang, Y. Lu, and H. Qiao, "Transformer-based spiking neural networks for multimodal audio-visual classification," *IEEE Transactions on Cognitive and Developmental Systems*, 2023.

[36] T. Chen, L. Wang, J. Li, S. Duan, and T. Huang, "Improving spiking neural network with frequency adaptation for image classification," *IEEE Transactions on Cognitive and Developmental Systems*, 2023.

[37] S. Geirnaert, T. Francart, and A. Bertrand, "Fast EEG-based decoding of the directional focus of auditory attention using common spatial patterns," *IEEE Transactions on Biomedical Engineering*, vol. 68, no. 5, pp. 1557–1568, 2020.

[38] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[39] S.-C. Liu, A. van Schaik, B. A. Minch, and T. Delbruck, "Asynchronous binaural spatial audition sensor with $2 \times 64 \times 4$ channel output," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 8, no. 4, pp. 453–464, 2013.

[40] E. M. Izhikevich, "Simple model of spiking neurons," *IEEE Transactions on Neural Networks*, vol. 14, no. 6, pp. 1569–1572, 2003.

[41] A. N. Burkitt, "A review of the integrate-and-fire neuron model: I. Homogeneous synaptic input," *Biological Cybernetics*, vol. 95, pp. 1–19, 2006.

[42] J. B. Aimone, Y. Ho, O. Parekh, C. A. Phillips, A. Pinar, W. Severa, and Y. Wang, "Provable advantages for graph algorithms in spiking neural networks," in *Proceedings of the 33rd ACM Symposium on Parallelism in Algorithms and Architectures*, 2021, pp. 35–47.

[43] J. Wu, C. Xu, X. Han, D. Zhou, M. Zhang, H. Li, and K. C. Tan, "Progressive tandem learning for pattern recognition with deep spiking neural networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 11, pp. 7824–7840, 2021.

[44] Y. Wang, M. Zhang, Y. Chen, and H. Qu, "Signed neuron with memory: Towards simple, accurate and high-efficient ann-snn conversion," in *International Joint Conference on Artificial Intelligence*, 2022.

[45] Y. Wu, L. Deng, G. Li, J. Zhu, and L. Shi, "Spatio-temporal backpropagation for training high-performance spiking neural networks," *Frontiers in neuroscience*, vol. 12, p. 331, 2018.

[46] N. Das, T. Francart, and A. Bertrand, "Auditory Attention Detection Dataset KULeuven," Aug. 2020, Version 1.1.0. [Online]. Available: https://doi.org/10.5281/zenodo.3997352

[47] S. A. Fuglsang, D. D. Wong, and J. Hjortkjær, "EEG and audio dataset for auditory attention decoding," Mar. 2018. [Online]. Available: https://doi.org/10.5281/zenodo.1199011

[48] P. Refaeilzadeh, L. Tang, and H. Liu, "Cross-validation." *Encyclopedia of Database Systems*, vol. 5, pp. 532–538, 2009.

[49] O.-Y. Kwon, M.-H. Lee, C. Guan, and S.-W. Lee, "Subject-independent brain–computer interfaces based on deep convolutional neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 10, pp. 3839–3852, 2020.

[50] S. Cai, P. Sun, T. Schultz, and H. Li, "Low-latency auditory spatial attention detection based on spectro-spatial features from EEG," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2021, pp. 5812–5815.

[51] V. L. Towle, J. Bolaños, D. Suarez, K. Tan, R. Grzeszczuk, D. N. Levin, R. Cakmur, S. A. Frank, and J.-P. Spire, "The spatial location of EEG electrodes: locating the best-fitting sphere relative to cortical anatomy," *Electroencephalography and Clinical Neurophysiology*, vol. 86, no. 1, pp. 1–6, 1993.

[52] M. Horowitz, "1.1 computing's energy problem (and what we can do about it)," in *2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC), IEEE*, 2014, pp. 10–14.

[53] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins *et al.*, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82–115, 2020.

[54] Y. Kim and P. Panda, "Visual explanations from spiking neural networks using inter-spike intervals," *Scientific Reports*, vol. 11, no. 1, pp. 1–14, 2021.

[55] J. Y. Shih, C. A. Atencio, and C. E. Schreiner, "Improved stimulus representation by short interspike intervals in primary auditory cortex," *Journal of Neurophysiology*, vol. 105, no. 4, pp. 1908–1917, 2011.

[56] A. L. Hodgkin and A. F. Huxley, "The components of membrane conductance in the giant axon of loligo," *The Journal of physiology*, vol. 116, no. 4, p. 473, 1952.