ABR-Attention: An Attention-Based Model for Precisely Localizing Auditory Brainstem Response

Junyu Ji, Xin Wang, *Member, IEEE*, Xiaobei Jing, Mingxing Zhu, Hongguang Pan, Desheng Jia, Chunrui Zhao, Xu Yong, Yangjie Xu, Guoru Zhao, *Member, IEEE*, Poly Z.H. Sun, *Member, IEEE*, Guanglin Li, *Senior Member, IEEE*, and Shixiong Chen

***Abstract*— Auditory Brainstem Response (ABR) is an evoked potential in the brainstem's neural centers in response to sound stimuli. Clinically, characteristic waves, especially Wave V latency, extracted from ABR can objectively indicate auditory loss and diagnose diseases. Several methods have been developed for the extraction of characteristic waves. To ensure the effectiveness of the method, most of the methods are time-consuming and rely on the heavy workloads of clinicians. To reduce the workload of clinicians, automated extraction methods have been developed. However, the above methods also have limitations. This study introduces a novel deep learning network for automatic extraction of Wave V latency, named ABR-Attention. ABR-Attention model includes a self-attention module, first and second-derivative attention module, and regressor module. Experiments are conducted on the accuracy with 10-fold cross-validation, the effects on different sound pressure levels (SPLs), the effects of different error scales and the effects of ablation. ABR-Attention shows efficacy in extracting Wave V latency of ABR, with an overall accuracy of 96.76 ± 0.41% and an error scale of 0.1ms, and provides a new solution for objective localization of ABR characteristic waves.**

***Index Terms*— Auditory brainstem response (ABR), deep learning network, ABR-attention.**

Manuscript received 3 January 2024; revised 8 August 2024; accepted 9 August 2024. Date of publication 19 August 2024; date of current version 5 September 2024. This work was supported in part by the National Key Research and Development Program of China under Grant 2022YFE0197500, in part by the National Natural Science Foundation of China under Grant 81927804 and Grant 62101538, in part by the Science and Technology Planning Project of Shenzhen under Grant JSGG20210713091808027 and Grant JCYJ20230807093819039, in part by the Postdoctoral Fellowship Program of China Postdoctoral Science Foundation (CPSF) under Grant GZC20241849, in part by China Postdoctoral Science Foundation under Grant 2024M753402, and in part by the Science and Technology Program of Guangdong Province under Grant 2022A0505090007. *(Junyu Ji, Xin Wang, and Xiaobei Jing contributed equally to this work.) (Corresponding authors: Hongguang Pan; Poly Z.H. Sun; Shixiong Chen.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Institutional Review Board (IRB) of Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, under Application No. SIAT-IRB-190615-H0352; and by Shenzhen Children's Hospital under Application No. 2022133.

Junyu Ji is with the CAS Key Laboratory of Human-Machine Intelligence-Synergy Systems, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, Guangdong 518055, China, also with the University of Chinese Academy of Sciences, Beijing 100049, China, and also with Guangdong-Hong Kong-Macao Joint Laboratory of Human-Machine Intelligence-Synergy Systems, Shenzhen, Guangdong 518055, China.

Xin Wang, Xiaobei Jing, Xu Yong, Guoru Zhao, and Guanglin Li are with the CAS Key Laboratory of Human-Machine Intelligence-Synergy Systems, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, Guangdong 518055, China, and also with Guangdong-Hong Kong-Macao Joint Laboratory of Human-Machine Intelligence-Synergy Systems, Shenzhen, Guangdong 518055, China.

Mingxing Zhu is with the School of Electronics and Information Engineering, Harbin Institute of Technology, Shenzhen 518055, China.

Hongguang Pan, Desheng Jia, and Chunrui Zhao are with the Department of Otolaryngology, Shenzhen Children's Hospital, Shenzhen 518033, China (e-mail: 1481717890@qq.com).

Yangjie Xu is with the Interdisciplinary Centre for Security, Reliability and Trust, University of Luxembourg, 1855 Luxembourg City, Luxembourg.

Poly Z.H. Sun is with the Department of Industrial Engineering, School of Mechanical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: zh.sun@sjtu.edu.cn).

Shixiong Chen is with the School of Medicine, The Chinese University of Hong Kong, Shenzhen, Guangdong 518172, China (e-mail: chenshixiong@cuhk.edu.cn).

Digital Object Identifier 10.1109/TNSRE.2024.3445936

I. Introduction

AUDITORY Brainstem Response (ABR) refers to the biophysical response that occurs in the brainstem part of the auditory pathway when external sound stimuli are applied to the human auditory system. This response can be recorded by placing electrodes on the scalp and using specific sound stimuli, such as clicks or short tonal pulses. It primarily occurs within 1-10 milliseconds after the stimulus. The response morphology varies depending on stimulus parameters (e.g., signal waveform, amplitude). Often only one of the five waves is extractable. In subjects with normal hearing, the ABR shows a very typical waveform, composed of several main waves, usually marked as Waves I to V. Each wave corresponds to the bioelectric activity of a specific part of the auditory pathway [1], [2]. For instance, Wave I is related to the activity near the cochlear nerve, while Waves III and V are related to the upstream brainstem neural structures. Waves I, III, and V have relatively large amplitudes, hence they are more widely used in clinical applications [3], [4], [5].

Since the wave characteristics of ABR can indicate human auditory pathway functions and diagnose neurological diseases, it has attracted great interest from the medical and biomedical engineering community [6], [7]. Moreover, due to its objectivity, non-invasiveness, and the ability to obtain results without the need for active responses, ABR has become the preferred tool for newborn hearing screening. Timely detection and intervention in newborn hearing loss are crucial for the development of language, cognition, and social skills in children. ABR can be used not only for the assessment of auditory function but also to help physicians determine specific lesion locations along the auditory pathway. For example, ABR in patients with acoustic neuroma shows abnormalities [8], hence it is used in the diagnosis of acoustic neuroma [9]. In cranial trauma detection, ABR also has its unique application. In many cases of cranial injury, ABR is often used as a tool to assess the function and integrity of the central nervous system. The abnormal ABR mainly manifests as abnormal waveforms, prolonged wave latency, low wave amplitude, and extended inter-wave periods [10]. In recent years, with the continuous progress and development of medical standards, ABR has been widely used in various auditory surgeries, such as cochlear implant surgery [11], acoustic neuroma surgery [12], and middle ear surgery. Meanwhile, ABR has been effectively used in the diagnosis of many diseases, such as Alzheimer's disease and geriatric schizophrenia identification, assessment of vertebrobasilar insufficiency, diagnosis of Parkinson's disease, hyperbilirubinemia, sudden deafness, central vestibular vertigo. Additionally, as an objective method of hearing assessment, ABR is not only widely used for newborn hearing screening but also applied in hearing detection and assessment of children with difficulties in subjective hearing tests and multiple disabilities, as well as in the objective detection of hearing loss in adults. Therefore, research on ABR has very important and positive significance for the development of neuroscience, life sciences, audiology, and clinical medicine.

Generally, Wave V in the conventional ABR has the largest amplitude among all positive peaks and occurs almost within 10 milliseconds after the sound stimulus, so the latency of Wave V has been extensively studied [13]. Until now, the extraction of the latency and amplitude of ABR waves has been completed by manually selecting the wave peaks and troughs. However, Zaitoun et al. show that the diagnostic results of ABR are currently related to the doctor's experience in interpreting waveforms, and significant differences often exist in the interpretation results of doctors with different experiences [14]. Even for experienced experts, manual marking requires a lot of time, especially when dealing with a large amount of data. More importantly, results obtained by different clinical doctors or the same clinical doctor in different states may also be different. In this way, the traditional methods are no longer satisfactory.

To solve this problem, researchers have conducted studies to avoid subjectivity. Elberling suggested correlating the individual auditory evoked potential signals with a standard response template to obtain an approximation of the latency of individual ABR [15]. Kneip and Gasser believe that this template can be considered an estimate of the common structure of individual responses, which is different from the traditional lateral response and is less affected by the smoothing effect of time variability [16]. However, this method only aligns the peaks of wave V. If the relative position between the individual wave crests changes, this affects the results. To solve this problem, researchers have proposed two different methods. The first one is the waveform template method proposed by Motsch [17]. This method uses a separate standard template to fit each characteristic wave of ABR. Unlike the response template method of Elberling et al., the template of this method is not obtained by measurement but is synthesized by functions and can be moved and scaled in time and amplitude, but the outcome is generally not ideal. Since the characteristic waves of ABR are mixed from a variety of electrophysiological activities from different sources, their intensity and timing may vary, thus becoming the limitation of the wave template method [18], [19], [20]. In comparison, the dynamic time-warping method proposed by Picton et al., like the response template method, is also a completely non-parametric method [21]. The difference is mainly that the dynamic time-warping method uses nonlinear time transformations, local stretching, or compression to align the characteristic waves, rather than using linear time to align individual responses. In addition, some researchers have used simpler methods to extract the characteristic waves of ABR, such as the derivative zero-crossing method [22], [23], [24]. This method extracts the peaks and troughs of ABR by deriving the ABR in time and finding the zero-crossing points of the derivatives. However, because the inherent noise in the ABR is disproportionately amplified during the derivation process, the selection of derivative zero-crossing points is usually limited to around the relevant wave peaks and troughs in the individual average response to deal with the noise problem, thus introducing selection bias.

Based on the aforementioned problems, researchers introduced artificial intelligence methods into this field. In audiology, the analysis and processing of ABR via machine learning and deep learning have also received widespread attention. Dogan adopted artificial neural networks to take the original ABR as input for ABR threshold detection [25]. The study showed that there is a close correlation between the clinical labels obtained by marking by clinical experts and the labels automatically generated by artificial neural networks. Chen et al. conducted a detailed study on the recognition of ABR features using deep learning and achieved good results [26]. However, only the analysis and testing of different structures of Long Short-Term Memory (LSTM) were carried out, and other network types such as Deep Neural Networks (DNN), Convolutional Neural Networks (CNN), and Transformer were not evaluated. Especially the Transformer model, an attention-based model, with the self-attention mechanism being a crucial component. This mechanism enables modeling of the relevance in input data, providing Transformers the ability to process long-range dependencies [27]. Thus, Transformers are widely applied in sequence data processing applications. Zheng et al. viewed semantic segmentation as a sequence and proposed the Segmentation Transformer (SETR) model to complete the sequence prediction task [28]. Zou et al. introduced a Transformer model for end-to-end object detection

[29]. Moreover, Transformers have also achieved significant results in processing biomedical sequence data. Guo et al. proposed a High-frequency oscillations detection framework based on Transformer for processing one-dimensional magnetoencephalography biomedical sequence data [30]. He et al. developed an EEG-Transformer model based on the traditional Transformer model, achieving good results in EEG signal classification [31]. Given the similarities between ABR and EEG signals, this study proposes a new model based on Transformer for processing ABR data and extracting the latency of Wave V, named ABR-Attention. ABR-Attention fully utilizes the attention mechanism of the Transformer, employing self-attention to consider the data's relevance, and integrating first and second-order derivative attention to make it more suitable for extracting the latency of ABR Wave V. The first and second-order derivatives of ABR data are significant in the extraction of Wave V latency [32]. Therefore, this study utilizes first and second-order derivative attention, fully considering the correlation between these derivatives and the ABR data itself.

II. Methods

A. Data Source

The ABR data was collected using the SmartEP system (*Intelligent Hearing, USA*) in an acoustic attenuation and electromagnetic shielding room. The study involved 1189 subjects who were subjected to click sound stimuli at intensities ranging from 10-100dB (in 10dB steps), yielding a total of 10841 ABR data. Out of these, 7585 ABR data entries with Wave V were selected. Among the 1189 subjects, there were 730 males and 459 females, aged 0-17 years (average 2.33 ± 2.91 years), including 464 with normal hearing and 725 with hearing abnormalities, which includes 141 conductive hearing loss ears, 980 sensorineural hearing loss ears, 36 mixed hearing loss ears, and 1107 normal ears. The summary is in TABLE I. Each ABR data point spans from −13.675ms to 11.900ms, with the 0ms mark indicating the moment of sound stimulus, encompassing 1024 sampling points at a sample frequency of 40 kHz. The wave V of ABR was marked by two experienced audiologists, which was recognized as golden latency. Finally, 90% of all ABR data was used as the training set and 10% as the test set. All experimental schemes were approved by the Institutional Review Board (IRB) of the Shenzhen Institutes of Advanced Technology, the Chinese Academy of Sciences (SIAT-IRB-190615-H0352), and the Shenzhen Children's Hospital (2022133).

B. Data Processing

In this study, two primary processes were applied to the original ABR data: 1) Cropping was done to eliminate interference from some data; 2) Normalization was applied to both the ABR data and the labels. The training set was also processed in two ways: 1) Dividing the entire training set into ten parts, with nine parts used as the actual training set and one part as the validation set. 2) Data augmentation, including scaling, noise injection, and cut-mix, was performed on the actual training set to increase the data volume.

TABLE I
Summary of the ABR Dataset

Gender	730 males, 459 females
Age	0-17 years (average 2.33±2.91 years)
Hearing	464 normal, 725 abnormal (hearing loss)
Etiology	141 conductive, 980 sensorineural, 36 mixed hearing loss, 1107 normal hearing

Cropping involved cutting the original ABR data −13.675∼11.9ms to 4∼11.5ms, selecting sampling points 708∼1008th from the 1024 range.

$$ABR\_all = [d_0, d_1, \cdots, d_{1023}] \tag{1}$$

$$ABR\_data = [d_{708}, d_{709}, \cdots, d_{1008}] \tag{2}$$

where *ABR_all* is the ABR data list of all sample points, *ABR_data* is the ABR data list after cropping, d_i is the ABR data sample point.

Normalization included normalizing the ABR data between 0∼1 based on maximum and minimum values. Since the latency of ABR Wave V under normal conditions is 5.69 ± 0.18ms [33], the study used 4.5ms and 11.5ms as the maximum and minimum values for label normalization.

$$\begin{cases} ABR\_data = \dfrac{ABR\_data - \min(ABR\_data)}{\max(ABR\_data) - \min(ABR\_data)} \\ Latency = \dfrac{golden\_latency - 5.5}{10.5 - 4.5} \end{cases} \tag{3}$$

where *ABR_data* is the ABR data list after cropping and normalization, *Latency* is the golden latency after normalization, and *golden_latency* is the latency of Wave V from the audiologist.

To evaluate the performance of the network, the 10 divided training sets were subjected to 10-fold cross-validation.

For data augmentation on the actual training set, cut-mix was first applied. The data after the cut-mix, combined with the original training set, formed a new training set. This new set underwent scaling at 0.6x and 0.8x and had white noise with an amplitude of 0.01 injected. All these steps formed the final training data.

To preserve the data-label correspondence, the cut-mix was performed by selecting 60 points before and after wave V for cutting and mixing, ensuring the mixed portions did not contain wave V. The labels for the new data remained the same as the original data.

$$\begin{cases} L\_p = \text{floor}(golden\_latency \times 40) \\ s\_p = \min(\max(0, L\_p0 - 60), \max(0, L\_p1 - 60)) \\ e\_p = \max(\min(300, L\_p0 + 60), \min(300, L\_p1 + 60)) \\ n\_d0 = \{d1[0 : s\_p], d0[s\_p : e\_p], d1[e\_p : 300]\} \\ n\_d1 = \{d0[0 : s\_p], d1[s\_p : e\_p], d0[e\_p : 300]\} \end{cases} \tag{4}$$

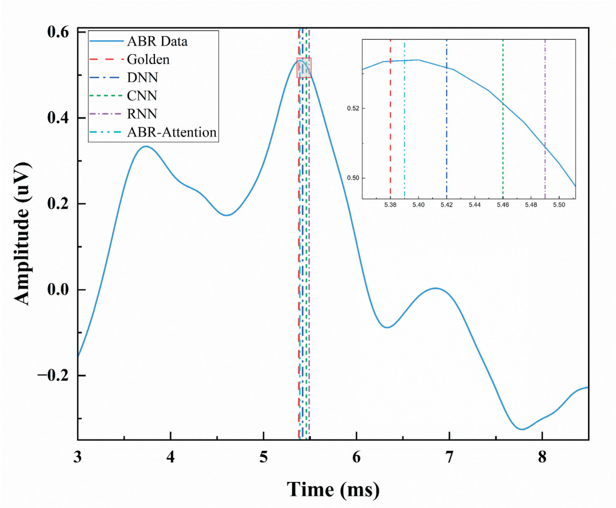

Fig. 1. An example of automatic ABR wave V latency extraction and localization by different deep learning models, with the golden line as the average of the manual markings by two independent experienced audiologists. Different colors of dotted lines represented the wave V localization outputs of different models.

where *golden_latency* is the latency of Wave V from the audiologist, *L_p* is the number of the latency sample point, *s_p* is the cut point before the latency point, *e_p* is the cut point after the latency point, *L_p0* is *L_p* of one ABR data and *L_p1* is of another one, *d0* is one ABR data list and *d1* is another, *n_d0* is new ABR data and *n_d1* is another new one.

Since the data was normalized, scaling was limited to 0.6x and 0.8x to keep the data within the 0∼1 range, without affecting the latency of the wave V.

$$\begin{cases} n\_d0 = d0 \times 0.6 \\ n\_d1 = d0 \times 0.8 \end{cases} \tag{5}$$

where *d0* is one ABR data list, *n_d0* is a new ABR data and *n_d1* is another new one.

To avoid impacting the overall trend of the data, white noise with an amplitude of 0.01 was injected. The labels for the data post-injection remained the same as those for the original data.

$$n\_d0 = d0 + \text{white\_noise}(0.01) \tag{6}$$

where *d0* is one ABR data list, and *n_d0* is a new ABR data. *white_noise(x)* will generate a white noise sequence with the same length as *d0* and an amplitude no greater than x.

C. Experimental Scheme

The objective of this study was to extract the latency of the ABR wave V, with ABR data as the input and the latency of the wave V as the output. We propose a novel model, namely ABR-Attention. And compared with other models, including DNN, CNN and RNN. Each model determined the latency of wave V through regression, deriving continuous wave V latency from ABR data. The illustration of the localization is shown in **Fig. 1**. Six experiments were designed in this study. The first was an accuracy experiment, where the accuracy of each fold of the four different models was tested using the test set. In this experiment, an error within 0.1ms was considered correct, while an error over 0.1ms was considered incorrect. The second was an experiment on different stimulus SPL, where the test set was divided into low ([20dB, 50dB] SPL), medium ((50dB, 70dB] SPL), and high ((70dB, 100dB] SPL) groups based on SPLs, to test the accuracy of different types of models, with the same error scale of 0.1ms. Third, we divided the test data into a normal hearing group and an abnormal hearing group for comparison, and here the error scale was also set to 0.1ms. The fourth was an error scale experiment, testing the accuracy of different models at various error scales (0.01∼0.2ms). Fifth, we analyzed the distribution of errors beyond 0.1ms to better observe the network's generalization ability and robustness. The final experiment was an ablation study, testing the accuracy of the ABR-Attention model as different network components were ablated, with an error scale of 0.1ms.

D. Network Structure

The main network structure in this study is based on the Self-Attention mechanism of the Transformer, widely used in single-sequence processing [34]. The mechanism focuses on different positions of a single sequence to calculate its attention, which applies to ABR data where different positions significantly impact wave V recognition.

In the identification of wave V, the first-order derivative and the second-order derivative also play an important roles [32]. The network pays attention not only to the ABR data itself but also to its first and second derivatives, using residual connections to retain the original ABR data information. The overall network architecture is illustrated in **Fig. 2**.

The attention mechanism structure, shown in **Fig. 2(a)**, inputs queries and keys of dimension and values of dimension. It computes the dot product of queries and keys, divided by, and then applies the softmax function to obtain the weights for the values. The formula is as follows:

$$\text{Attention}(Q, K, V) = \text{soft}\max(\frac{QK^T}{\sqrt{d_k}})V \tag{7}$$

where Q is the Query, K represents the Key, V means the Value, and d_i is the length of the tensor.

In the traditional Self-Attention, queries, keys, and values are derived from the original data (ABR waveform data) through a linear layer. In the improved network, attention to the data itself and its first and second derivatives are included. This part modifies the keys in the Attention mechanism to be derived from the first and second derivatives. The formulas are as follows:

$$\begin{cases} Q_i = \text{liner\_q}_i(ABR\_data) & i = 0, 1, 2 \\ K_0 = \text{liner\_}k_0(ABR\_data) & \\ K_1 = \text{liner\_}k_1(first\_derivative) & \\ K_2 = \text{liner\_}k_2(second\_derivative) & \\ V_i = \text{liner\_v}_i(ABR\_data) & i = 0, 1, 2 \end{cases} \tag{8}$$

where Q_i is the Query of the i-th attention, K_i is the Key of the i-th attention, V_i is the Value of the i-th attention i = *1, 2, 3*, *ABR_data* is the ABR waveform data, *first_derivative* is

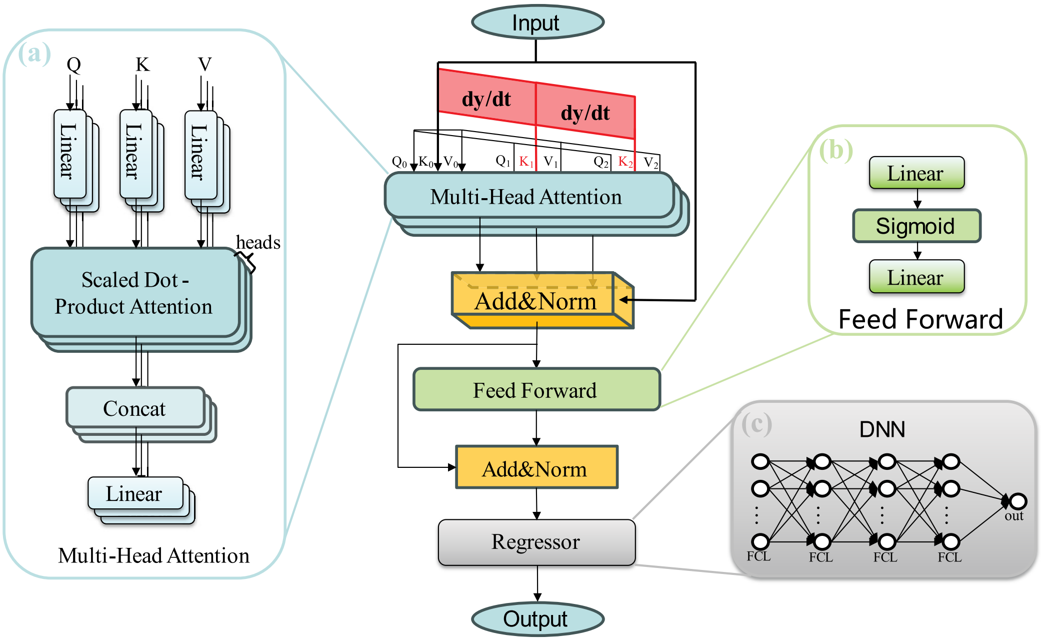

Fig. 2. The proposed ABR-attention deep learning structure to automatically extract wave V latency of ABR signals. (a) the structure of multi-head attention mechanism, (b) the structure of feed forward procedure, (c) the structure of the regressor to extract ABR latency using the output of the transformer model.

the first derivative of the *ABR_data*, *second_derivative* is the second derivative of the *ABR_data*.

The individual attention components are combined into a Multi-Head Attention module, enabling the model to focus on more sub-space information. The data itself, along with its first and second derivatives, are processed through the Multi-Head Attention module. After processing, the data is combined using Add&Norm for residual connection. The formulas are as follows:

$$\begin{cases} \text{Multi-Head}(Q, K, V) = \text{Concat}(h_0, h_1, \ldots, h_n)W^o \\ h_i = \text{Attention}(Q, K, V) \\ MH_i = \text{Multi-Head}_\text{i}(Q_i, K_i, V_i) \\ \qquad i = 0, 1, 2 \\ Add\_data = ABR\_data + MH_0 + MH_1 + MH_2 \\ Add\&Norm_0(Add\_data) = \text{Normalize}(Add\_data) \end{cases} \tag{9}$$

where MH_0 is the result of multi-head self-attention, MH_1 is the result of multi-head first derivative attention, and MH_2 is the result of multi-head second derivative attention.

Subsequently, the data passes through a Feed Forward layer, which is shown in **Fig. 2(b)** The Feed Forward layer contains two fully connected layers, and the activation function of the first fully connected layer is Sigmoid. The output also passes through the Add&Norm layer for residual connection. The formulas are as follows:

$$\begin{cases} \text{FeedForward}(x) = \text{sigmoid}(xW_0 + b_0)W_1 + b_1 \\ Add\&Norm_1(x) = \text{Normalize}(x + \text{FeedForward}(x)) \end{cases} \tag{10}$$

where W_i are the weights, and b_i are the biases.

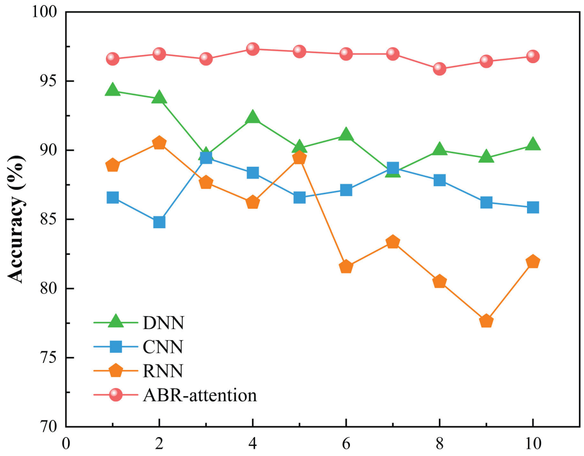

Fig. 3. The accuracy of the 10-fold cross-validation.

Finally, the data undergoes regression calculation through a regressor to determine the latency of ABR Wave V. The regressor is a small-scale DNN network composed of three fully connected layers, with node counts of 2400, 3600, and 1, respectively, which is shown in **Fig. 2(c)**.

E. Statistical Analysis

The accuracies of the 10-fold tests for different deep learning models were statistically analyzed under different experimental conditions. For each stimulus level, the accuracies of different models were statistically compared using one-way ANOVA and pairwise comparisons were conducted among the models if significant difference was found. Meanwhile, a one-way ANOVA was also conducted on different stimulus levels for each model to examine the level effects.

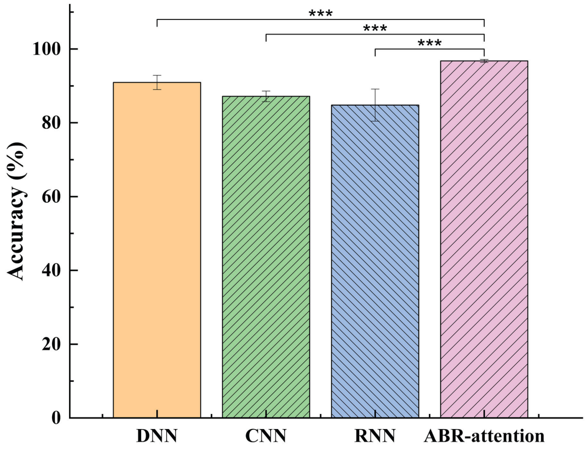

Fig. 4. The statistical comparison of the accuracy in the ABR wave V latency extraction among different deep learning models (DNN, CNN, RNN and the proposed ABR-attention; *** represented p < 0.001).

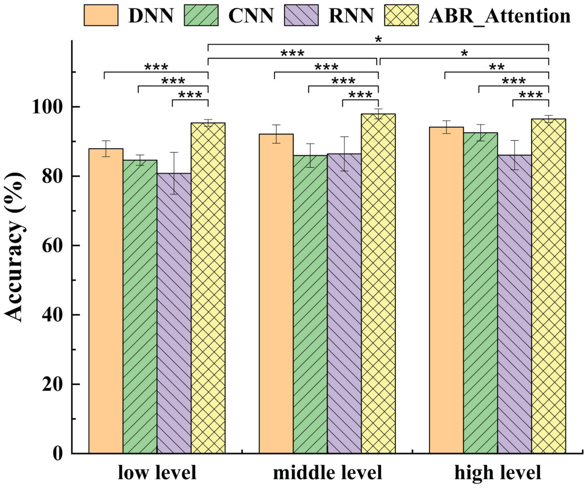

Fig. 5. The statistical comparison of the performance of different models in extracting ABR Wave V latency under low, middle and high stimulus levels (* represented p < 0.05, ** represented p < 0.01, *** represented p < 0.001).

Then the means and standard deviations of the performances of different deep learning models were compared and plotted with statistical results at different significance levels, with * represented p < 0.05, ** represented p < 0.01, *** represented p < 0.001.

III. Result

A. Performance on Accuracy

This section employs a 10-fold cross-validation method to test the accuracy of four different deep-learning models. These models include DNN, CNN, RNN, and the ABR-Attention model.

The results are shown in **Fig. 3**, where ABR-attention has the highest 10-fold accuracy among all networks, and the accuracy of each fold is above 95%, and the fluctuation between each fold is minimal. The accuracy of DNN is around 90%, but its fluctuation is relatively large. The accuracy of CNN and RNN is almost below 90%, and the accuracy of RNN fluctuates the most between each fold. We performed a statistical analysis on the 10-fold accuracy. As shown in **Fig. 4**, ABR-attention has the highest accuracy of 96.76 ± 0.41% and is significantly different from other models.

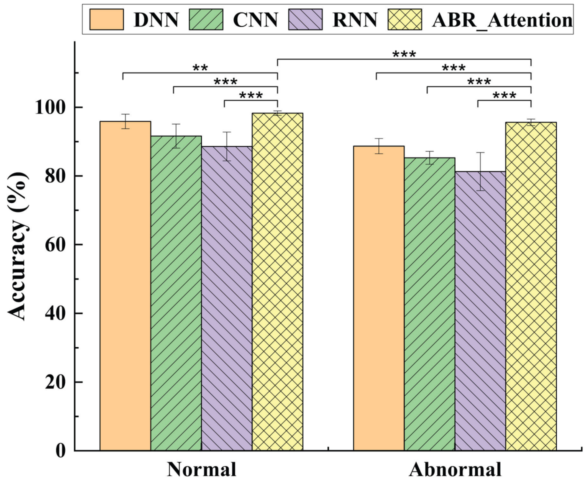

Fig. 6. Accuracy of normal and abnormal hearing (** represented p < 0.01. *** represented p < 0.001).

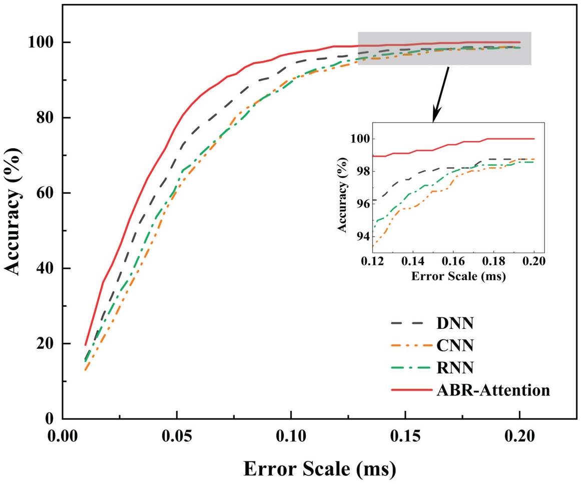

Fig. 7. Result of experiments with different error scales.

B. Experiments With Different SPLs

In this part, the accuracy of four different neural network models (DNN, CNN, RNN, and ABR-Attention) are compared across three levels of SPL: low, middle, and high.

As illustrated in **Fig. 5**, the ABR-Attention model outperforms the other models across all SPL categories consistently and is significantly different from other models. At mid-level SPL, ABR-Attention has the highest accuracy of 97.94 ± 1.43% and is significantly different from low SPL and high SPL.

C. Performance in Normal and Abnormal Groups

The accuracy in ABR Wave V latency extraction of the different models was also systemically compared for normal and abnormal hearing groups and the results were shown in **Fig. 6**.

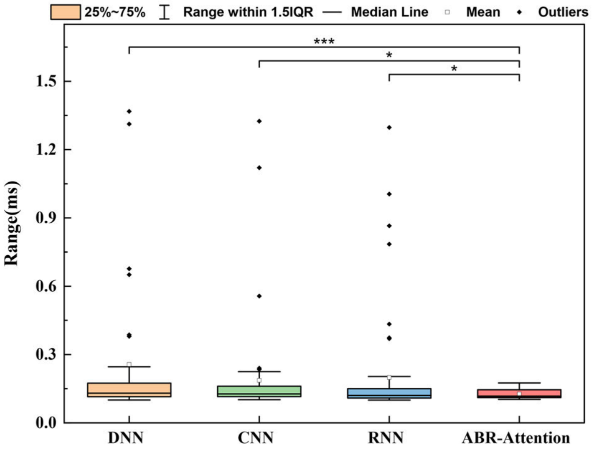

Fig. 8. The distribution of the extracted latencies of different models for the incorrectly localized trials according to the 0.1ms criteria (* represented $p < 0.05$, *** represented $p < 0.001$).

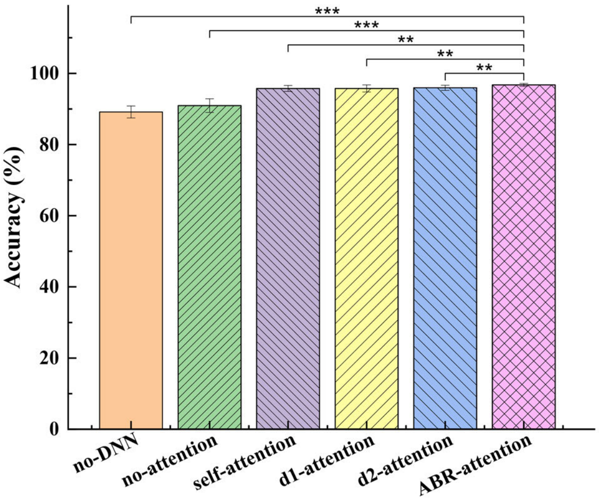

Fig. 9. The comparison of the accuracy performance for the proposed ABR-attention model for different ablation experiments (** represented $p < 0.01$, *** represented $p < 0.001$).

As shown in the figure, our proposed ABR-attention model showed significantly higher accuracies than other models, for both the normal group and the abnormal group. Meanwhile, the accuracy of the normal group was also significantly higher than that of the abnormal group, regardless of the type of network. The accuracy of ABR-attention in the normal group could reach 98.29 ± 0.66%, while the accuracy for the abnormal group was 95.60 ± 0.95%.

D. Experiments With Different Error Scales

This section assesses the accuracy of four distinct models across various error scales ranging from 0.01 to 0.20 milliseconds.

The results, as depicted in Fig. 7, the accuracy of all models increases with the increase of the allowed error scale. ABR-attention has the best accuracy at all error scales. When the error scale is 12ms, the accuracy exceeds 99%, and when the error scale is 0.17ms, the accuracy reaches 100%.

E. Analysis of Errors Beyond the 0.1ms

This section analyzes the errors beyond the 0.1ms for different network models. It primarily examines number of the pints, the median, concentration, mean, and outliers. The results are shown in Fig. 8. As shown in the figure, the medians of all four methods are in the lower range, with ABR-Attention having the lowest median. RNN exhibits the largest data dispersion, indicated by a taller box and multiple outliers, while ABR-Attention shows the smallest dispersion, with a shorter box and fewer outliers. The means of the four methods are not significantly different and are close to the medians within the boxes. Additionally, the RNN method has the most outliers, indicating extreme values in certain cases. And ABR-attention is significantly different from other models.

F. Ablation Experiments

This section details ablation experiments for the ABR-Attention model, where different components of the model are removed to assess their impact on accuracy. This includes the removal of the DNN regressor (no-DNN), the omission of the attention mechanism (no-attention), the exclusion of all enhancements (self-attention), the removal of second-order derivative enhancements (d1-attention), the removal of first-order derivative enhancements (d2-attention), and the complete ABR-Attention model (ABR-attention). The details are in TABLE II, and the results are presented in Fig. 9 and Fig. 10.

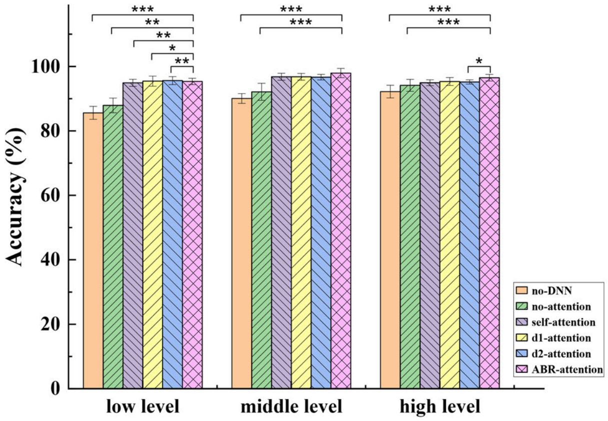

Fig. 10. The performance comparison of the proposed ABR-attention model for different ablation experiments under low, middle and high stimulus levels (* represented $p < 0.05$, ** represented $p < 0.01$, *** represented $p < 0.001$).

Fig. 9 shows the accuracy of the model. It shows that if there is no DNN regressor or no attention mechanism, the accuracy rate will be reduced more. After the attention mechanism is introduced, the accuracy rates of self-attention, d1-attention, and d2-attention are not much different. ABR-attention has the highest accuracy and is significantly different from other models.

Fig. 10 shows the results of the ablation experiments with different SPL. At low-level SPL, ABR-attention and other ablation models have significant differences, and the accuracy of all models with attention mechanisms is similar.

TABLE II
STRUCTURE AND RESULT OF ABLATION EXPERIMENTS

Net Type	**Self Attention**	**1st -derivative Attention**	**2nd-derivation Attention**	**Regressor**	**Accuracy (%)**
no-DNN	√	√	√	×	**89.14±1.69**
no-attention	×	×	×	√	**90.93±1.92**
self-attention	√	×	×	√	**95.76±0.84**
d1-attention	√	√	×	√	**95.76±1.00**
d2-attention	√	×	√	√	**95.94±0.72**
ABR-attention	√	√	√	√	**96.76±0.41**

TABLE III
COMPARISON OF ACCURACY OF DIFFERENT METHODS

Methods	**Accuracy (%)**	
	Error scale (0.1ms)	**Error scale (0.2ms)**
[26]	85.46	92.91
[32]	Not mentioned	96-98
DNN	88.37-94.28	98.75
CNN	84.79-89.45	98.75
RNN	77.64-90.52	98.57
ABR-attention	95.89-97.32	100.00

ABR-attention has the highest accuracy at middle-level SPL and is only significantly different from no-DNN and no-attention. At high-level SPL, ABR-attention has the highest accuracy and is significantly different from no-DNN, no-attention and d2-attention.

IV. Discussion

Auditory Brainstem Responses (ABRs) are produced when the synchronous neural fiber encodes sound. To assist medical personnel in more conveniently obtaining clinical parameters of ABR, this paper proposes a novel deep learning network, ABR-Attention, for extracting the position of wave V of ABR. Four experiments are designed to demonstrate the network's efficacy.

A. Performance on Accuracy

This study employed the K-Fold cross-validation method to verify Performance on Accuracy, K-Fold cross-validation is a method used to validate neural networks to determine predictability [35]. Due to limited data volume, we chose K=10 for cross-validation to increase the amount of training data. For ease of result comparison, we fitted the 10-fold results, and the fit curve visually indicates the level of network accuracy.

Based on the aforementioned methodology, the proposed ABR-Attention model was compared with three other models (DNN, CNN, and RNN) to study the performance of extracting the latency of ABR Wave V. The extraction of Wave V latency is a typical regression task. RNNs excel at processing sequential data as they can retain previous information, aiding in better understanding the sequence context [36]. However, although ABR is sequential, each input is a complete ABR, and there's no temporal relationship between two inputs, nor is there a need to retain previous information. Therefore, RNNs do not perform well in extracting ABR Wave V latency. CNNs were initially designed for multi-dimensional array data [37], and since ABR data is one-dimensional, CNNs do not leverage their strengths in this task, leading to subpar performance. One reason for CNN's poor performance is the potential overfitting issue when handling low-dimensional data [38]. In contrast, DNNs perform relatively well with ABR data, but they only consider the data itself, limited by the quality of the ABR data. ABR-Attention, which considers not only the ABR data itself but also the relationships between ABR data point-to-point (i.e., Self-Attention), the correlation between ABR data and its first-order derivative (i.e., first-order derivative Attention), and the correlation between ABR and its second-order derivative (i.e., second-order derivative Attention), achieves the highest accuracy.

Due to the introduction of the derivative attention mechanism, ABR-attention has better performance than ordinary deep learning methods [26] in extracting Wave V latency. At the same time, due to the combination of deep learning and derivatives, it also has better performance than traditional derivative methods [32]. As shown in the TABLE III.

B. Experiments With Different SPLs

In this experiment, the ABR-Attention model exhibited the best performance across all SPL levels, and there is a significant difference from other models. It is widely acknowledged that ABR waveforms obtained at high-level SPLs should have more pronounced Wave V, making them easier to identify. However, the highest accuracy was achieved at mid-level SPLs, which may be attributed to the greater quantity of data available at mid-level SPLs compared to high-level SPLs. Despite this, the smaller error lines in accuracy at high-level SPL recognition also precisely demonstrate that more pronounced Wave V leads to more stable recognition outcomes. There are significant differences in all levels of SPL for ABR-attention, which also illustrates that ABR-attention has different capabilities in processing different SPL data.

C. Performance in Normal and Abnormal Groups

As we expected, ABR-attention has the highest accuracy in the normal group and is significantly different from the abnormal group. This maybe because the lesions of the auditory system cause some implicit information in ABR to be missing.

D. Experiments With Different Error Scales

ABR-attention showed high accuracy at all error scales, which is consistent with our expected results. When the error scale is 0.12ms, the accuracy can reach 99%. Within the allowable error, ABR-attention is fully capable of clinical tasks.

E. Analysis of Errors Beyond the 0.1ms

ABR-Attention has the lowest median and smallest data dispersion, indicating more stable performance across various scenarios with fewer extreme values. This stability is crucial for practical applications as it ensures more consistent results. The minimal data dispersion and few outliers suggest that the ABR-Attention model is robust in handling different data distributions or noise. Investigating the model's robustness and adaptability to various data environments can further optimize and promote this method.

F. Ablation Experiments

The purpose of this experiment is to demonstrate the necessity of different components. As hypothesized, regressors play an integral role in research. We also found that introducing a single attention mechanism can improve performance significantly. ABR-attention enhances its error correction capabilities by focusing on three aspects: itself, first-order derivatives, and second-order derivatives, thereby achieving higher efficiency when considering all attention mechanisms. There are significant differences between ABR-attention and other ablated models, which shows that each component is meaningful.

In the results across different SPL levels, we found consistency with the findings in Chapter IV-B, which showed the highest accuracy at mid-level SPL due to the influence of data volume. However, in the results at high-level SPL, we observed that both relative accuracy and error lines were better, indirectly suggesting that ABR-Attention's dependency on data volume is relatively reduced.

G. Generalization Ability of the Proposed Model

Generalization ability is a key indicator for evaluating the performance of neural networks, and it is also a key factor in determining whether the model can be widely used in clinical applications. For the novel multi-head attention mechanism proposed by this study, it not only paid attention to the pattern of ABR temporal waveform but also simultaneously attended to other useful features (such as the first-order and second-order derivatives of the ABR time waveform) to improve the generalization ability of the transformer model. The extra features introduced by this study were closely related to the unique patterns of the Wave V peak and played an important role in localizing the Wave V position. Our results showed that the proposed method by this study significantly outperformed all other studies, with an accuracy of 95.89-97.32% in the 0.1ms error tolerance scale. The results suggested our proposed multi-head attention method can improve the generalization ability of ABR Wave V latency extraction tasks, by introducing more clinically useful information during the feature attention stage in the improved transformer model. It is also noteworthy that more diverse and larger datasets should be included to further improve the generalization ability [39], since only the data from children under the age of 17 were involved in this study. In the future, more ABR data from more diverse populations (such as the adult population and newborn neonates) should be included to improve the generalization ability and to validate the model's applicability.

V. Conclusion

Given that current methods in automatically extracting the ABR wave V locations showed quite limited accuracy and performance, the main innovation of our study is to propose a new deep learning model based on a novel multi-head attention mechanism employed in Transformer models. For the novel multi-head attention mechanism proposed by this study, it not only paid attention to the pattern of ABR temporal waveform (called self-attention mechanism as most Transformer studies employed), but also simultaneously attended to other useful features (such as the first-order and second-order derivatives of the ABR time waveform) to improve the performance of the transformer model. The extra features introduced by this study were closely related to the unique patterns of Wave V peak and played an important role in localizing the Wave V position [32]. Our results showed that the proposed method by this study significantly outperformed all other studies, with an accuracy of 95.89-97.32% in the 0.1ms error tolerance scale. The results suggested our proposed multi-head attention method can significantly improve the accuracy of ABR Wave V latency extraction tasks, by introducing more clinically useful information during the feature attention stage in the improved transformer model.

References

[1] J. A. Sisneros, A. N. Popper, A. D. Hawkins, and R. R. Fay, "Chapter 130 auditory evoked potential audiograms compared with behavioral audiograms in aquatic animals," *Adv. Experim. Med. Biol.*, vol. 875, pp. 1049–1056, Aug. 2016.

[2] J. J. Eggermont, "Auditory brainstem response," in *Handbook of Clinical Neurology*, vol. 160. Amsterdam, The Netherlands: Elsevier, Jul. 2019, pp. 451–464.

[3] J. D. Lewis, J. Kopun, S. T. Neely, K. K. Schmid, and M. P. Gorga, "Tone-burst auditory brainstem response wave V latencies in normal-hearing and hearing-impaired ears," *J. Acoust. Soc. Amer.*, vol. 138, no. 5, pp. 3210–3219, Nov. 2015.

[4] A. L. Smit et al., "Automated auditory brainstem response in preterm newborns with histological chorioamnionitis," *J. Maternal-Fetal Neonatal Med.*, vol. 28, no. 15, pp. 1864–1869, Oct. 2015.

[5] M. Zaitoun, S. Cumming, A. Purcell, and K. O'brien, "Inter and intra-reader variability in the threshold estimation of auditory brainstem response (ABR) results," *Hearing, Balance Commun.*, vol. 14, no. 1, pp. 59–63, Jan. 2016.

[6] M. S. Robinette, C. D. Bauch, W. O. Olsen, S. G. Harner, and C. W. Beatty, "Use of TEOAE, ABR, and acoustic reflex measures to assess auditory function patients with acoustic neuroma," *Amer. J. Audiology*, vol. 1, no. 4, pp. 66–72, Nov. 1992, doi: 10.1044/1059-0889.0104.66.

[7] E. Kimura et al., "Effect of shock wave power spectrum on the inner ear pathophysiology in blast-induced hearing loss," *Sci. Rep.*, vol. 11, no. 1, p. 14704, Jul. 19, 2021, doi: 10.1038/S41598-021-94080-0.

[8] M. Montaguti, C. Bergonzoni, M. A. Zanetti, and A. Rinaldi Ceroni, "Comparative evaluation of ABR abnormalities in patients with and without neurinoma of VIII cranial nerve," *Acta Otorhinolaryngol Ital*, vol. 27, no. 2, pp. 68–72, Apr. 2007.

[9] W. A. Selters and D. E. Brackmann, "Acoustic tumor detection with brain stem electric response audiometry," *Arch. Otolaryngology-Head Neck Surgery*, vol. 103, no. 4, pp. 181–187, Apr. 1977.

[10] H. Wang et al., "High frequency of AIFM1 variants and phenotype progression of auditory neuropathy in a Chinese population," *Neural Plasticity*, vol. 2020, pp. 1–12, Jul. 2020, doi: 10.1155/2020/5625768.

[11] H. Yang, J. Tang, K. Cao, X. Zhu, Y. Wang, and T. Pan, "The intraoperative application of neural response telemetry with the nucleus CI24M cochlear implant," *Zhonghua Er Bi Yan Hou Ke Za Zhi*, vol. 36, no. 5, pp. 352–356, 2001.

[12] D. Han, L. Yu, S. Yang, and L. Yu, "Hearing protection during the operation of acoustic neurinoma," *Chin. J. Otology*, vol. 3, pp. 8–174, Dec. 2004.

[13] K. C. Backer, A. S. Kessler, L. A. Lawyer, D. P. Corina, and L. M. Miller, "A novel EEG paradigm to simultaneously and rapidly assess the functioning of auditory and visual pathways," *J. Neurophysiology*, vol. 122, no. 4, pp. 1312–1329, Oct. 2019, doi: 10.1152/JN.00868.2018.

[14] M. Zaitoun, S. Cumming, A. Purcell, and K. O'Brien, "The impact of clinical history on the threshold estimation of auditory brainstem response results for infants," *J. Speech, Lang., Hearing Res.*, vol. 60, no. 3, pp. 725–731, Mar. 2017.

[15] C. Elberling, "Auditory electrophysiology: The use of templates and cross correlation functions in the analysis of brain stem potentials," *Scandin. Audiology*, vol. 8, no. 3, pp. 187–190, Jan. 1979.

[16] A. Kneip and T. Gasser, "Statistical tools to analyze data representing a sample of curves," *Ann. Statist.*, vol. 20, no. 3, pp. 1266–1305, Sep. 1992.

[17] J.-F. Motsch, "La dynamique temporelle du tronc cerebral: Recueil, extraction et analyse optimale des potentiels evoques auditifs du tronc cerebral," Ph.D. thesis, France, 1987. [Online]. Available: https://theses.fr/1987PA120013.

[18] L. J. Achor and A. Starr, "Auditory brain stem responses in the cat. I. Intracranial and extracranial recordings," *Electroencephalogr. Clin. Neurophysiology*, vol. 48, no. 2, pp. 154–173, Feb. 1980.

[19] D. L. Jewett and J. S. Williston, "Auditory-evoked far fields averaged from the scalp of humans," *Brain*, vol. 94, no. 4, pp. 681–696, 1971.

[20] D. L. Jewett, M. N. Romano, and J. S. Williston, "Human auditory evoked potentials: Possible brain stem components detected on the scalp," *Science*, vol. 167, no. 3924, pp. 1517–1518, Mar. 1970.

[21] T. Picton, M. Hunt, R. Mowrey, R. Rodriguez, and J. Maru, "Evaluation of brain-stem auditory evoked potentials using dynamic time warping," *Electroencephalogr. Clin. Neurophysiology Evoked Potentials Sect.*, vol. 71, no. 3, pp. 212–225, May 1988.

[22] R. Schaette and D. McAlpine, "Tinnitus with a normal audiogram: Physiological evidence for hidden hearing loss and computational model," *J. Neurosci.*, vol. 31, no. 38, pp. 13452–13457, Sep. 2011.

[23] H. Guest, K. J. Munro, G. Prendergast, R. E. Millman, and C. J. Plack, "Impaired speech perception in noise with a normal audiogram: No evidence for cochlear synaptopathy and no relation to lifetime noise exposure," *Hearing Res.*, vol. 364, pp. 142–151, Jul. 2018.

[24] H. Guest, K. J. Munro, and C. J. Plack, "Tinnitus with a normal audiogram: Role of high-frequency sensitivity and reanalysis of brainstem-response measures to avoid audiometric over-matching," *Hearing Res.*, vol. 356, pp. 116–117, Dec. 2017.

[25] D. Alpsan, "Classification of auditory brainstem responses by human experts and backipropagation neural networks," in *Proc. Annu. Int. Conf.*, 1991, pp. 1425–1426.

[26] C. Chen et al., "Automatic recognition of auditory brainstem response characteristic waveform based on bidirectional long short-term memory," *Frontiers Med.*, vol. 7, Jan. 2021, Art. no. 613708.

[27] T. Wang et al., "O-net: A novel framework with deep fusion of CNN and transformer for simultaneous segmentation and classification," *Frontiers Neurosci.*, vol. 16, Jun. 2022, Art. no. 876065, doi: 10.3389/FNINS.2022.876065.

[28] S. Zheng et al., "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Aug. 6890, pp. 6881–6890.

[29] C. Zou et al., "End-to-end human object interaction detection with hoi transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Sep. 2021, pp. 11825–11834.

[30] J. Guo et al., "Transformer-based high-frequency oscillation signal detection on magnetoencephalography from epileptic patients," *Frontiers Mol. Biosciences*, vol. 9, Mar. 2022, Art. no. 822810, doi: 10.3389/FMOLB.2022.822810.

[31] Y. He et al., "Classification of attention deficit/hyperactivity disorder based on EEG signals using a EEG-transformer model," *J. Neural Eng.*, vol. 20, no. 5, Oct. 2023, Art. no. 056013.

[32] A. P. Bradley and W. J. Wilson, "Automated analysis of the auditory brainstem response using derivative estimation wavelets," *Audiology Neurotology*, vol. 10, no. 1, pp. 6–21, 2005, doi: 10.1159/000081544.

[33] J. Krizman, E. Skoe, and N. Kraus, "Stimulus rate and subcortical auditory processing of speech," *Audiology Neurotology*, vol. 15, no. 5, pp. 332–342, 2010, doi: 10.1159/000289572.

[34] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–14.

[35] J. Molina, A. Laroche, J.-V. Richard, A.-S. Schuller, and C. Rolando, "Neural networks are promising tools for the prediction of the viscosity of unsaturated polyester resins," *Frontiers Chem.*, vol. 7, p. 375, May 2019, doi: 10.3389/FCHEM.2019.00375.

[36] Y. Wang et al., "Causal discovery in radiographic markers of knee osteoarthritis and prediction for knee osteoarthritis severity with attention–long short-term memory," *Frontiers Public Health*, vol. 8, Dec. 2020, Art. no. 604654, doi: 10.3389/FPUBH.2020.604654.

[37] F. Alharbi and A. Vakanski, "Machine learning methods for cancer classification using gene expression data: A review," *Bioengineering*, vol. 10, no. 2, p. 173, Jan. 28, 2023, doi: 10.3390/BIOENGINEERING10020173.

[38] H. Li et al., "An interpretable computer-aided diagnosis method for periodontitis from panoramic radiographs," *Frontiers Physiol.*, vol. 12, Jun. 2021, Art. no. 655556, doi: 10.3389/FPHYS.2021.655556.

[39] Z. Tan et al., "Fast anther dehiscence status recognition system established by deep learning to screen heat tolerant cotton," *Plant Methods*, vol. 18, no. 1, p. 53, Apr. 21, 2022, doi: 10.1186/S13007-022-00884-0.