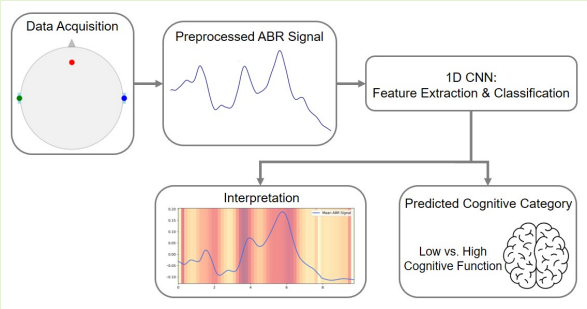
Predicting Cognitive Performance from Three-Electrode Auditory Brainstem Responses Using Convolutional Neural Networks

Mobina Malekifar, *Graduate Student Member, IEEE*, Yasmeen Hamza, Ye Yang, Hung Cao, *Senior Member, IEEE*, and Fan-Gang Zeng, *Fellow, IEEE*

***Abstract*— Age-related cognitive decline is a growing global concern, motivating the search for non-invasive, accessible biomarkers to support early detection and monitoring. Click-evoked auditory brainstem responses (ABRs), collected in routine clinical settings, offer a promising signal source. Building on prior evidence that ABRs relate to cognitive function, this study investigates whether raw human ABR waveforms can predict cognitive performance using deep learning without manual peak measurements. We used a dataset from 118 adults spanning a broad range of cognitive abilities, pairing each click-evoked ABR with a cognitive score. A one-dimensional convolutional neural network (CNN) was trained to learn time-series patterns directly from the raw signal. Model performance was evaluated using two, five, and ten-fold cross-validation and compared against traditional wave V metrics and a randomized-input baseline. After adjusting scores for age, the CNN achieved a mean area under the receiver operating characteristic curve of 0.77 ± 0.06, outperforming all benchmarks. To interpret model decisions, Gradient-weighted Class Activation Mapping (Grad-CAM) was applied. Three key latency windows were identified: 1.8 to 2.3 ms, 3.2 to 4.0 ms, and 4.9 to 6.5 ms. These correspond to canonical ABR waves I, III, and V, supporting the physiological relevance of the learned features and highlighting Wave III as a previously underutilized marker. Although limited to click stimuli from a single recording system, this work demonstrates that a CNN can extract meaningful features from raw ABRs collected using only three electrodes and predict cognitive status more accurately than traditional methods using the same dataset.**



***Index Terms*— Auditory Brainstem Response (ABR), Cognition, Deep Learning, Convolutional Neural Network (CNN).**

I. Introduction

COGNITIVE health is a vital aspect of overall well-being, particularly in the context of neurodegenerative diseases such as Alzheimer's Disease (AD) and related dementias, which currently affect over 55 million individuals worldwide. These conditions also carry a significant economic burden, with estimated annual costs reaching approximately 1.3 trillion USD [1]. Identifying a robust, accessible, and non-invasive biomarker for cognitive function could greatly enhance the ability to track cognitive aging, inform targeted interventions, and support early detection and treatment strategies for AD and similar disorders [2], [3].

While current biomarkers have improved our understanding of cognitive decline, they often rely on fluid analysis or imaging methods, such as magnetic resonance imaging (MRI) and positron emission tomography (PET), to detect structural and metabolic brain changes [4]–[6]. These approaches may be invasive, costly, or unsuitable for frequent or long-term monitoring in large populations [6]. Recent research highlights electrophysiological signals as a promising, objective tool for tracking cognitive decline, overcoming the limitations of existing biomarkers. By linking brain activity during tasks like language and memory retrieval to clinical symptoms, these measures could serve as a non-invasive, cost-effective assessment method in routine evaluations [7], [8].

This work was supported in part by the NSF-UKRI/BBSRC-/BIO #2321840 (F.G.Z) and the NSF EFRI #2422412 (H.C.). (Corresponding authors: Hung Cao and Fan-Gang Zeng.)

Mobina Malekifar is with the Center for Hearing Research and the Department of Electrical Engineering and Computer Science, University of California at Irvine, Irvine, CA 92697 USA (e-mail: nmalekif@uci.edu).

Yasmeen Hamza is with the Institute of Sound and Vibration Research, School of Engineering, University of Southampton, Southampton, UK (e-mail: Y.Hamza@soton.ac.uk).

Ye Yang is with the Center for Hearing Research and the Department of Biomedical Engineering, University of California at Irvine, Irvine, CA 92697 USA (e-mail: yey27@uci.edu).

Hung Cao is with the Department of Electrical Engineering and Computer Science, the Department of Biomedical Engineering, and the Department of Computer Science, University of California at Irvine, Irvine, CA 92697 USA (e-mail: hungcao@uci.edu).

Fan-Gang Zeng is with the Center for Hearing Research, the Department of Otolaryngology-Head and Neck Surgery, the Department of Anatomy and Neurobiology, the Department of Biomedical Engineering, and the Department of Cognitive Sciences, University of California, Irvine, CA 92697 USA (e-mail: fzeng@uci.edu).

Hearing-related brain signals are especially relevant in this context. A strong link has been observed between age-related hearing loss and cognitive decline [9]. In fact, hearing impairment is considered the most significant modifiable midlife risk factor for developing dementia [10]. Furthermore, the use of hearing aids has been associated with a reduced risk of dementia among older adults who are vulnerable to cognitive impairment [11]. Building on the findings of Gray et al. [12], who reported that slower temporal processing in the auditory brainstem is linked to reduced cognitive performance in aging rhesus macaques, our previous work [13] aimed to investigate whether a similar relationship exists in humans. To explore this, we recorded auditory brainstem responses (ABRs) using a 60 dB sensation level (SL), 51 Hz click train, and assessed cognitive performance across ten distinct tasks in a group of 118 adults.

However, in that work, ABR analysis relied on manual peak identification, a process that requires substantial expertise and time. It remains prone to human error, especially with atypical waveforms, and lacks standardization for certain conditions [14]–[16]. Similarly, traditional computational methods have improved efficiency but still rely on hand-crafted features, limiting their ability to capture abnormal or complex patterns. These limitations highlight the need for more flexible approaches. Deep learning addresses this gap by learning directly from raw signals, offering improved accuracy, objectivity, and scalability in ABR analysis.

Some studies have explored this direction, applying neural networks to ABR data for waveform classification and hearing assessment, highlighting the potential of learning-based methods in auditory research [17]. For instance, a feedforward network introduced in [18] successfully distinguished normal from abnormal ABRs but relied on manually extracted features from the time and frequency domains.

In another study, Seha et al. [19] explored the use of one-dimensional convolutional neural networks (1D CNNs) to classify signals for biometric authentication, demonstrating the potential of deep learning to learn discriminative features from auditory signals for subject identification. Similarly, McKearney and MacKinnon [20] explored several neural network architectures, including multilayer perceptrons (MLPs), recurrent neural networks (RNNs) with Long Short-Term Memory (LSTM) units, bidirectional LSTMs (BiLSTMs), and a deep convolutional neural network (CNN), to classify paired ABR waveforms into three clinically relevant categories: clear response, inconclusive, and response absent. They found that a deep CNN performed best for the multi-class clinical classification of ABRs, achieving 92.9% accuracy. Building on this trend, Chen et al. [15] developed a BiLSTM-based model to automatically identify characteristic waves I, III, and V in ABR recordings.

A more recent investigation by Ma et al. [21] introduced a preprocessing pipeline and deep learning framework to classify hearing loss using ABR graph images. The study utilized 10,000 ABR samples evenly divided between normal hearing and hearing loss cases. A CNN-based model was then trained on the processed images to classify hearing status. The latest advancement in this field, presented by Liang et al. [22], introduced a CNN-BiLSTM-Attention model to classify ABR waveforms across individuals with varying ages and hearing levels, demonstrating the effectiveness of integrating temporal modeling with clinical context for robust ABR interpretation.

While existing studies have made significant progress in automated ABR analysis for waveform classification, threshold estimation, and wave identification, few have explored its potential for cognitive assessment. Building on the success of these end-to-end learning strategies, this study addresses this gap by developing a 1D CNN framework to predict cognitive performance directly from raw ABR waveforms. This approach moves beyond our previous work, which depended on manually extracted features [13], by investigating whether a deep learning model can achieve comparable predictive performance through automated analysis of the full waveform. We also visualize which temporal regions of the signal contribute most to the model's predictions to enhance interpretability.

To our knowledge, this is the first study to apply deep learning to ABRs for cognitive score prediction. The remainder of this paper is organized as follows. First, we describe the dataset and our proposed model (Section II). We then present the evaluation results (Section III) and discuss their context and limitations (Section IV). Finally, Section V concludes with a summary of our contributions and future research directions.

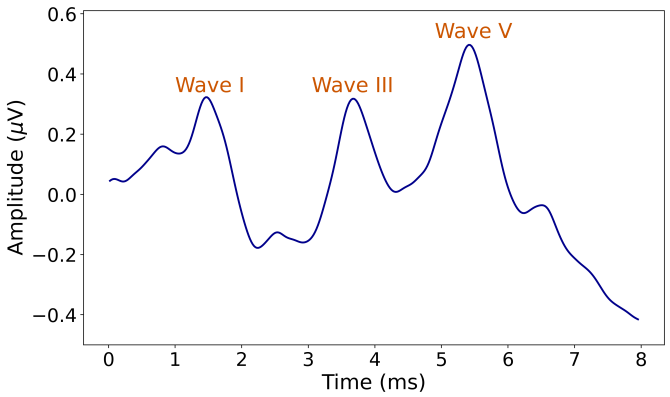


Fig. 1: Sample ABR waveform from a normal-hearing subject, with wave V marked at approximately 5.41 ms.

II. Materials and Methods

A. Experimental Data

In this study, we used the dataset previously analyzed in Hamza et al. [13]. Data collection procedures are summarized here. The dataset included auditory and cognitive data from 130 adult participants. Twelve individuals were excluded due to incomplete cognitive testing, inability to complete ABR recordings, or the presence of conductive hearing loss. This resulted in a final sample of 118 participants (66 female) spanning ages 18–92 years, divided into three age groups: young adults (18–30 years; n = 26), middle-aged adults (31–59 years; n = 26), and older adults (60–92 years; n = 66). Hearing thresholds ranged from -5 to 70 dB hearing level (HL). A total of 74 participants had normal hearing, defined as a pure-tone average (PTA) $<$ 20 dB HL, including all individuals in the

young group. The remaining 44 participants had hearing loss (PTA $\geq$ 20 dB HL), which was observed only in the middle-aged and older groups. Otoscopic screening was conducted to rule out visible external and middle ear abnormalities in all participants. The study protocol was approved by the University of California, Irvine Institutional Review Board, and all participants provided written informed consent.

Each participant underwent hearing and cognitive assessments. Pure-tone audiometry measured hearing thresholds across standard octave frequencies from 0.125 to 8 kHz. The better ear, identified by the average threshold at 0.5, 1, 2, and 4 kHz, was selected for all subsequent auditory testing.

Cognitive performance was assessed using a battery of eight standardized tests that produced ten outcome measures. These tests evaluated domains such as memory, executive function, attention, processing speed, and visuospatial ability. Tasks included word learning and recall, trail-making, symbol-digit substitution, verbal fluency, recognition memory, visual discrimination, and spatial matching. Lists and descriptions of the full set of standardized cognitive tests are included in the supplementary document.

ABRs were recorded from the better ear using an insert earphone system with gold tiptrodes and standard surface electrode placement. The tiptrode served as the inverting electrode at the test ear and the ground at the contralateral ear, while the non-inverting electrode was positioned at the high forehead (Fz). Electrode impedance was kept below 3 kΩ to ensure high-quality signal acquisition. The electrode configuration is shown in Fig. 2.

A 100 µs click stimulus was presented in alternating polarity at a rate of 51.33 Hz. Participants were asked to indicate whether they could detect this specific stimulus to determine the lowest audible level for each individual. The ABR recording stimulus was then delivered at 60 dB above each participant's click detection threshold. Each ABR recording used an epoch window of 10.66 ms, with each window containing 512 data points. Recordings were conducted in a soundproof booth with participants seated comfortably.

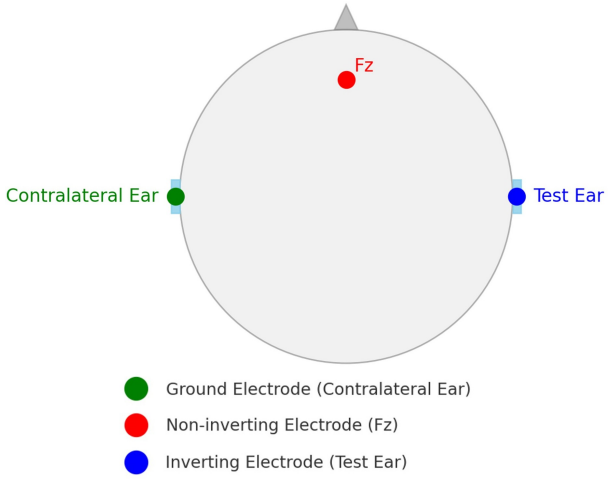


Fig. 2: Electrode placement setup for ABR data acquisition from the right ear.

B. Preprocessing

Each of the ten cognitive scores from eight tests was z-score normalized and adjusted so higher scores indicated better performance. A composite z-score was then computed by averaging across all measures. ABR signals were bandpass filtered from 100 to 3000 Hz, and epochs exceeding 23.8 µV were excluded. A minimum of 2000 artifact-free sweeps were averaged per recording.

The original dataset contained ABR signals recorded at multiple stimulation frequencies and intensity levels. To remain consistent with the previous study, only waveforms corresponding to 51 Hz stimulation and 60 dB SL were used as model input. Pre-stimulus samples were removed to ensure alignment, and only post-stimulus portions were retained. Due to varying pre-stimulus lengths, waveform sizes differed and were zero-padded within batches to match the longest signal. This padding does not affect classification performance, as zero regions typically do not activate convolutional units, reducing unnecessary weight updates and improving training efficiency [23].

Although input normalization (e.g., scaling to [-1, 1]) is common in deep learning to aid convergence [24], early experiments showed better performance without it. As a result, ABR signals were fed into the network using their original amplitudes. One possible explanation is that normalization may suppress meaningful inter-subject variability in signal amplitude. This is particularly relevant given that previous findings have shown a relationship between wave V amplitude and cognitive performance [13].

Each waveform was converted to a PyTorch tensor of shape [1, T], with T varying by subject, to match the 1D CNN input format. For binary classification, cognitive z-scores were transformed into percentile ranks (1-99%) by ranking subjects within the sample distribution and scaling the ranks to percentiles. At each 1% increment, binary labels were assigned: subjects at or above the threshold were labeled as high performers (1), and those below as low performers (0). This enabled model training and evaluation across the full range of cognitive performance thresholds.

C. Model Architecture

The model in this paper is a 1D CNN that processes variable-length ABR waveforms ([1, T]) and predicts the probability of a subject's cognitive performance exceeding a given percentile threshold, enabling binary classification of cognitive performance. The 1D CNN architecture is well-suited for time-series data, as it learns local and hierarchical patterns directly from raw input. Unlike traditional methods with hand-crafted features, 1D CNNs integrate feature extraction and prediction within a unified framework, enhancing efficiency and scalability in biomedical signal processing [25]. CNNs are also robust to distortions like phase shifts and amplitude scaling, extracting stable, noise-resilient features even under low signal-to-noise conditions [26].

As illustrated in Fig. 3, the architecture consists of four sequential convolutional blocks. The first convolutional layer uses 16 filters with a kernel size of 5 and padding of 2,

followed by batch normalization, a rectified linear unit (ReLU) activation, and a max pooling layer. ReLU outputs the input if it is positive and zero otherwise, introducing non-linearity into the network. All convolutional layers use a stride of 1, while max pooling operations downsample the feature maps with a kernel size and stride of 2. This pattern is repeated with 32 and 64 filters in the second and third layers, respectively, both using a kernel size of 5 (with padding 2 for the second layer and 3 for the third). The fourth block employs 128 filters with a kernel size of 3 and padding of 1, concluding with adaptive average pooling to produce a fixed-length output regardless of input duration. This feature is critical for handling variability in signal duration across subjects.

The resulting output is then flattened and passed to a fully connected classification head consisting of three dense layers with 256, 128, and 64 units, respectively. Each layer is followed by a ReLU activation and dropout regularization with dropout rates of 0.4, 0.3, and 0.2. The final layer is a single linear output unit producing a logit, with a sigmoid activation implicitly applied via the binary cross-entropy loss function during training.

D. Model Interpretability

Deep learning models are often viewed as black boxes, raising concerns about interpretability in clinical settings. To address this, we used gradient-weighted class activation mapping (Grad-CAM) [27] to highlight regions of the ABR signal most influential in predicting cognitive performance. Grad-CAM, a generalization of Class Activation Mapping (CAM) [28], is well-suited for CNN architectures that include fully connected layers after convolutional blocks. It identifies salient regions by computing a weighted combination of the feature maps from the last convolutional layer. In the proposed architecture, it is applied to the convolutional layer that outputs 128 feature maps just before the adaptive average pooling layer. This layer was chosen as it preserves sufficient temporal resolution while capturing high-level abstract features relevant to the model's prediction.

At inference time, Grad-CAM estimates the contribution of each feature map channel k to a target class c, which in this context corresponds to whether a subject's cognitive score exceeds a given percentile threshold. This is done by averaging the gradient of the output logit y^c with respect to the activation map A^k across the temporal dimension of the feature map:

$$\alpha_k^c = \frac{1}{Z} \sum_j \frac{\partial y^c}{\partial A_j^k} \tag{1}$$

where Z denotes the number of temporal positions. The resulting weights α_k^c are then used in a weighted summation of the activation maps:

$$L_{\text{Grad-CAM}}^c = \text{ReLU}\left(\sum_k \alpha_k^c A^k \right) \tag{2}$$

The ReLU activation ensures that only features positively associated with the output class are emphasized in the final attribution map. This results in a time-resolved saliency representation that highlights the portions of the ABR waveform most relevant to the model's classification decision at each cognitive performance threshold.

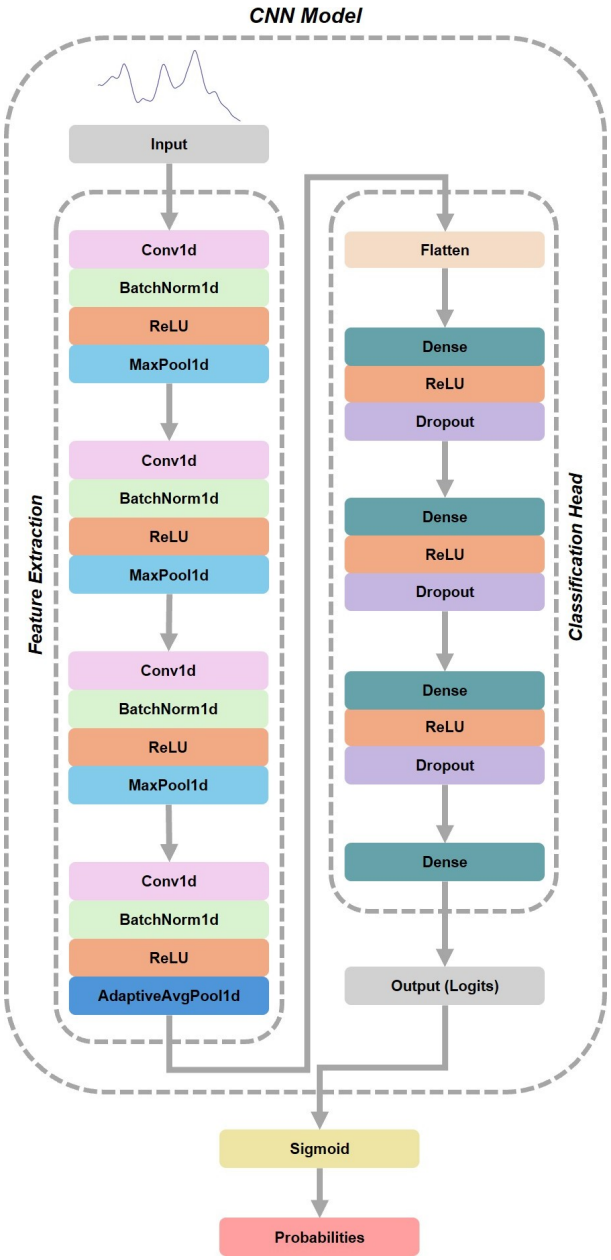


Fig. 3: Architecture of the 1D CNN used to classify cognitive performance from raw ABR signals. The model consists of a feature extraction module with four convolutional blocks, each followed by normalization, activation, and pooling layers. The resulting feature map is flattened before entering the classification head, which includes fully connected layers with dropout. The network produces a single logit that is passed through a sigmoid function to yield a probability for threshold-based binary classification of cognitive z-scores.

E. Training and Evaluation

The model was trained using the Adam optimizer [29] with a learning rate of 0.001 and a batch size of 16. Binary cross-entropy loss with logits was used to optimize predictions for

binary classification at each percentile threshold. This function combines a sigmoid activation and binary cross-entropy loss into a single, numerically stable operation.

Given a batch of predictions $x = \{x_1, \ldots, x_N\}$ paired with corresponding binary labels $y = \{y_1, \ldots, y_N\}$, where N is the batch size, the per-sample loss is defined as:

$$\ell_n = -\left[y_n \cdot \log(\sigma(x_n)) + (1 - y_n) \cdot \log(1 - \sigma(x_n))\right], \quad (3)$$

where $\sigma(x_n) = \frac{1}{1+\exp(-x_n)}$ is the sigmoid activation of the raw model output (logit). The final loss used for backpropagation is computed as the average over all individual sample losses in the batch:

$$\ell(x, y) = \frac{1}{N} \sum_{n=1}^{N} \ell_n \quad (4)$$

To assess model robustness, we used K-fold cross-validation [30] with k=2, 5, and 10. Data were randomly shuffled and split subject-wise to prevent overlap between training and test sets. In each iteration, one fold was used for testing and the remaining k-1 for training, repeating k times. Metrics were averaged across folds for each cognitive threshold. If a fold contained only one class (e.g., at 1% or 99% thresholds), evaluation metrics could not be calculated, and such folds were excluded from averaging at that threshold.

To enable comparison with the previous study, model performance was evaluated using area under the curve (AUC), defined as the area under the receiver operating characteristic (ROC) curve. AUC was computed across all cognitive percentile thresholds. ROC curves were generated at three representative thresholds (10%, 50%, 90%) using predicted probabilities, with true positive rate (TPR) and false positive rate (FPR) calculated at various decision thresholds using class labels derived from each representative percentile. They are calculated as

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (5)$$

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \quad (6)$$

where TP (true positives), FN (false negatives), FP (false positives), and TN (true negatives) are defined based on the binary classification at each threshold. TPR measures the proportion of correctly identified positive cases (high cognitive performers), while FPR reflects the proportion of negative cases (low performers) incorrectly classified as positive.

III. Results

A. Model Performance Across Cognitive Thresholds

Fig. 4 shows the 5-fold cross-validated performance of the CNN model in distinguishing between higher and lower cognitive scores across the full percentile range (1% to 99%). As presented, not only did the CNN model perform consistently above the chance level, but it also demonstrated similar or better performance than the wave V latency and amplitude predictor, particularly in the mid-percentile range.

To quantitatively compare the methods, Table I reports the mean ± standard deviation of AUC scores across cognitive thresholds for each approach, along with p-values from paired t-tests comparing the CNN model with wave V amplitude, wave V latency, and chance performance, using 2-, 5-, and 10-fold cross-validation, consistent with the fold settings in the previous study. Across all fold settings, the CNN model significantly outperformed chance and demonstrated superior performance compared to both wave V amplitude and latency in predicting age-adjusted scores.

Fig. 5 presents ROC curves at three representative cognitive thresholds: 10%, 50%, and 90%. These thresholds were selected to illustrate model performance across the full spectrum of cognitive percentiles and to compare the CNN-based approach with traditional methods. Consistent with earlier findings, the CNN model trained on age-adjusted cognitive scores outperformed both wave V amplitude and latency at lower and mid-range thresholds (10% and 50%). However, at the 90% threshold, where amplitude is more informative for identifying higher cognitive performers, wave V amplitude showed performance nearly equal to that of the CNN model.

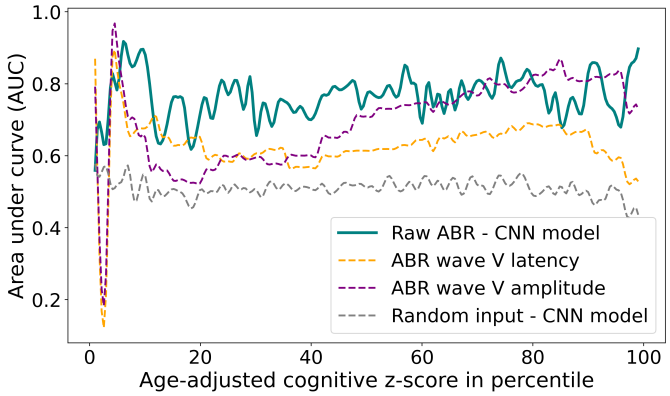


Fig. 4: AUC as a function of cognitive performance percentile, comparing results based on different input biomarkers. The teal line represents raw ABR signals analyzed by the CNN model, while the purple and orange dashed lines correspond to ABR wave V amplitude and latency, respectively. The gray dashed line indicates chance performance, obtained by feeding randomly generated input signals to the CNN model. All results are averaged over 5-fold cross-validation.

B. Visualization of Key Temporal Regions

Fig. 6 illustrates Grad-CAM attribution results based on 5-fold cross-validation analysis. Panel 6(a) shows the mean Grad-CAM attribution scores ± standard deviation across cognitive percentile thresholds for age-adjusted cognitive scores. Panel 6(b) presents heatmaps of normalized mean Grad-CAM attribution scores overlaid on the average ABR waveform across all subjects. Because ABR signals varied in length across subjects, shorter signals were padded by repeating their final data point to match the length of the longest signal, ensuring consistency prior to averaging. Darker red regions indicate higher attribution values, highlighting temporal segments of the ABR waveform that were most influential in the CNN model's predictions.

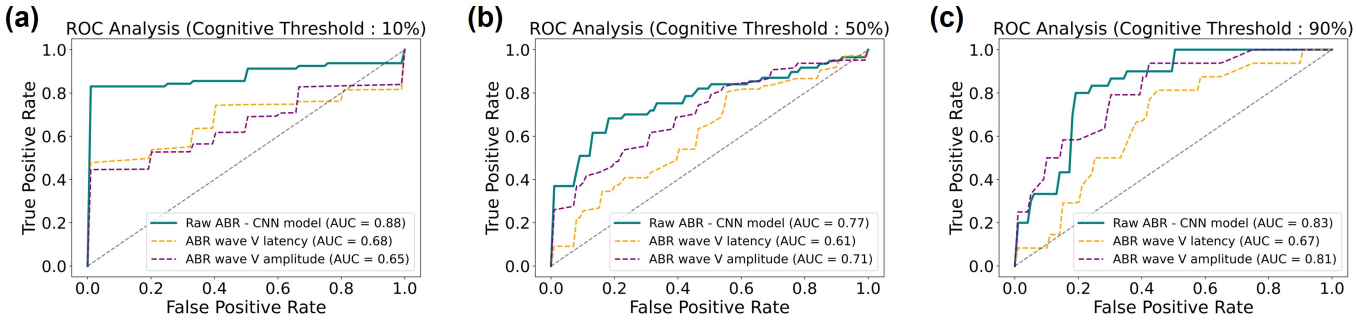


Fig. 5: ROC analysis at cognitive thresholds of 10%, 50%, and 90%, comparing the CNN model (solid teal) with ABR wave V amplitude (purple dashed) and latency (orange dashed). AUC values are shown for each method. The gray diagonal indicates chance-level performance. Results are based on 5-fold cross-validation using age-adjusted scores.

TABLE I: Comparison of AUC values (mean ± standard deviation) from the proposed CNN model and previously used methods, based on 2-, 5-, and 10-fold cross-validation. Paired t-test p-values indicate whether differences in predictive performance are statistically significant.

Number of Folds	Raw ABR - CNN Model	ABR Wave V Amplitude	ABR Wave V Latency	Random Input - CNN Model
2	0.77±0.06	0.70±0.10	0.60±0.09	0.51±0.02
Paired t-test	–	t = 6.65, $p < 0.001$	t = 15.59, $p < 0.001$	t = 39.99, $p < 0.001$
5	0.77±0.06	0.69±0.12	0.63±0.08	0.51±0.02
Paired t-test	–	t = 6.36, $p < 0.001$	t = 14.49, $p < 0.001$	t = 36.66, $p < 0.001$
10	0.77±0.08	0.72±0.11	0.65±0.07	0.51±0.02
Paired t-test	–	t = 4.56, $p < 0.001$	t = 11.18, $p < 0.001$	t = 30.26, $p < 0.001$

Based on this figure, the model consistently focuses on two primary temporal regions: approximately 3.2-4.0 ms and, to a lesser extent, the narrow 1.8-2.3 ms region and the wider 4.9-6.5 ms region, reflecting a smaller yet meaningful contribution to the model's predictions. Smaller spikes in attribution appear near the signal's start and end; however, these are brief (less than 0.2 ms), likely have minimal influence on the model's decisions, and may lack clear physiological relevance.

IV. Discussion

The primary goal of this study was to evaluate whether a one-dimensional CNN trained on raw ABR waveforms could match or exceed the performance of traditional methods based on Wave V latency and amplitude in predicting cognitive scores. As shown in Table I, the CNN results were significantly better than those obtained using Wave V latency and amplitude. Across 2-, 5-, and 10-fold cross-validation, the CNN consistently outperformed both chance and conventional biomarkers when predicting age-adjusted cognitive scores, achieving a mean AUC of 0.77 ± 0.06, compared with 0.65 ± 0.07 and 0.72 ± 0.11 for Wave V latency and amplitude, respectively. The fact that performance remained consistent across multiple cross-validation folds suggests that these results are robust, rather than being artifacts of a specific data split. This points to the existence of stable, generalizable features within the ABR waveform that the model is successfully capturing.

As illustrated in Fig. 4 and Fig. 5, the CNN's greatest advantage over traditional ABR markers occurs in the mid-percentile range of cognitive scores, where it consistently outperforms both Wave V latency and amplitude. The ROC analyses further indicate that the CNN achieves superior sensitivity–specificity trade-offs at these intermediate thresholds. This is particularly notable because the mid-range represents the majority of the population. Accordingly, the CNN's ability to capture subtle waveform variations in this range may be critical for the early detection of cognitive decline. In contrast, performance at higher cognitive thresholds appears to be driven primarily by Wave V amplitude, suggesting that robust neural synchrony remains the strongest predictor [31].

While the earlier study [13] explored both adjusted and unadjusted scores, we focused on adjusted scores, as performance on unadjusted scores declined, particularly at the highest percentiles where sample sizes were small, although still above chance. A key reason for this choice is that age adjustment removes a major confounder: chronological aging, which independently affects ABR features even in healthy individuals [32]. This allows the CNN to focus on waveform features reflecting cognitive status rather than age-related variance and prevents the model from relying on demographic shortcuts, encouraging learning of true biophysical cues in the ABR waveform instead of noncausal age effects.

Additionally, age adjustment leads to a more homogeneous training distribution by reducing variability in the ABR-cognition relationship. This supports the network in converging on robust, generalizable feature sets, particularly at higher cognitive percentiles where subject counts are limited. As the number of folds increases, class imbalance worsens and model performance declines, which underscores the importance of

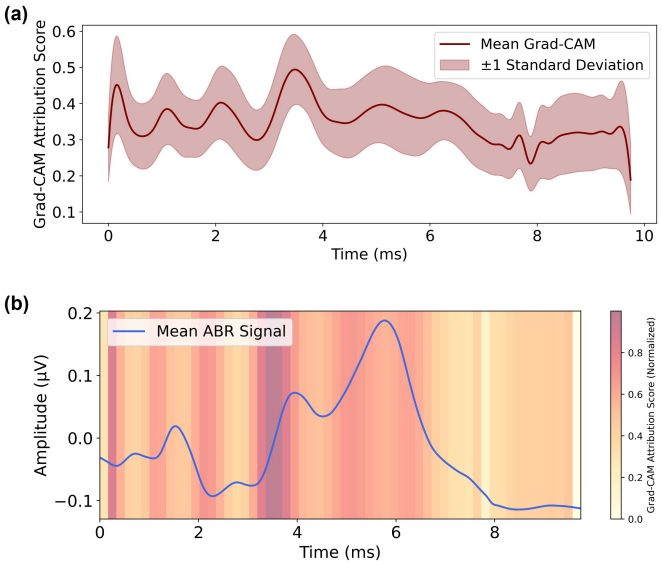


Fig. 6: Visualization of Grad-CAM attribution scores averaged across all subjects, 5-fold cross-validation, and cognitive performance thresholds.(a) Mean Grad-CAM scores (± standard deviation) over time for age-adjusted thresholds, plotted up to the longest input signal. (b) Normalized Grad-CAM overlay on the average ABR waveform, highlighting temporally important regions; darker red indicates higher importance.

addressing imbalance when predicting rare outcomes [33].

In terms of interpretability, Grad-CAM analysis showed that the CNN's predictions depend primarily on three temporal windows: 1.8 to 2.3 ms, 3.2 to 4.0 ms, and 4.9 to 6.5 ms. These intervals align with established ABR landmarks. In healthy adults, click-evoked ABRs typically show Wave I around 1 to 2 ms, Wave III around 3 to 4 ms, and Wave V around 5 to 6 ms [34]. Our previous work [13] estimated Wave I between 1.3 and 2.3 ms and Wave V between 5.1 and 6.4 ms. In Fig.6, the CNN's strongest attributions appear first in the Wave III window, then in the Wave V window, and finally in the Wave I window, aside from brief edge spikes that lack physiological relevance. The darkest red regions in the Grad-CAM heatmap in Fig. 6(b) coincide with the three main peaks of the averaged ABR waveform. A side-by-side comparison with the labeled peaks in Fig. 1 confirms that these attributions correspond precisely to Waves I, III, and V in a typical normal-hearing subject.

Although earlier studies emphasized Wave V as the primary subcortical marker linked to cognition, our results show that the model focused most on the Wave III window. This suggests that Waves I, III, and V all convey meaningful cognitive information and that future studies should examine Wave III more closely when developing ABR-based cognitive prediction methods.

Integrating deep learning with ABR protocols helps overcome a key limitation of ABR-based cognitive screening: reliance on expert-driven waveform interpretation. Although ABR measurements are objective, low-cost, and passive, their broader adoption has been limited by the need for manual analysis. The proposed CNN automates feature extraction with strong performance, enabling scalable ABR-based screening in consumer devices, including hearables and wearables. Compared with behavioral tests that depend on attention and are affected by repeated testing, an ABR-based approach can be applied repeatedly over time, including during low-engagement states such as rest or sleep.

Despite promising results, this study has some important limitations. First, the dataset is relatively small and contains few ABR observations associated with each cognitive score, especially in the highest cognitive percentile ranges. Deep learning models generally require large, diverse samples to learn stable and generalizable feature representations; limited data increase the risk of overfitting and performance instability [35], [36]. This scarcity likely contributed to the weaker results for unadjusted scores and restricts overall generalizability.

Second, all recordings were acquired with a single evoked-potential system under controlled laboratory conditions. Variability in clinical hardware, electrode placement, and ambient noise may challenge the model in real-world settings. Third, we considered only click-evoked ABRs, so it remains unclear whether the model would perform equally well on tone-burst or speech-evoked responses. In addition, similar to prior work based on this dataset, the analysis did not account for other potentially influential factors, including sex, head size, education level, and socioeconomic status. A larger sample would be required to disentangle the effects of these variables on the relationship between ABR measures and cognition.

Finally, we evaluated only a single network architecture. Although a 1D CNN is straightforward and computationally efficient, models that integrate recurrent layers, attention mechanisms, or transformer blocks (recent work shows transformers can capture broad time and frequency structure in raw electroencephalography signals [37]) may extract multiscale features from ABR waveforms more effectively. Moreover, we restricted our evaluation to AUC to enable direct comparison with the earlier study. Including additional metrics, such as the area under the precision-recall curve, calibration error, and decision curve analysis, would offer a better assessment of clinical utility, especially in the presence of class imbalance.

Future work should broaden both the data foundation and the modeling strategy for ABR-based cognitive prediction. Larger multicenter datasets that span different ages, hearing profiles, and recording hardware will improve generalizability and allow systematic tests of cross-device robustness. On the modeling side, future studies should evaluate hybrid networks that couple convolutional encoders with recurrent layers, attention modules, or transformer blocks. It will also be important to test tone-burst and speech-evoked ABRs to determine whether diverse acoustic cues improve prediction.

Additionally, based on our findings in [13], the Word Learning and Animal Fluency tests (see the supplementary file for more information about these tests) were associated with both ABR Wave V latency and amplitude. This suggests that these cognitive domains may contribute more strongly to the relationship between ABR features and overall cognitive performance and should be a focus of continued research. Combining manually derived features such as wave latencies

and amplitudes with deep-learned representations may further boost accuracy, especially when data remain limited. In light of our Grad-CAM results, future models should explicitly incorporate information from Wave III alongside Wave I and Wave V to capture the full range of cognitively relevant brainstem activity.

V. Conclusion

In this study, we demonstrated that a 1D CNN can predict cognitive performance directly from raw ABR waveforms, outperforming traditional Wave V biomarkers. Interpretability analysis confirmed the model learned biologically plausible features, focusing on canonical waves I, III, and V, and highlighted a potentially overlooked role for Wave III in cognitive assessment. While this work provides a strong proof-of-concept, future studies should validate these findings on multi-site data and with more complex stimuli before clinical application. Ultimately, this deep learning approach lays the groundwork for transforming standard ABR recordings into a rapid and objective method for monitoring cognitive health.

Conflicts of Interest

F.G.Z. owns stock in Axonics, Neocortix, Nurotron, Syntiant and Velox Biosystems. The other authors declare no competing interests.

References

[1] W. H. Organization *et al.*, "Global action plan on the public health response to dementia 2017–2025," in *Global action plan on the public health response to dementia 2017–2025*, 2017.

[2] T. Hedden, A. P. Schultz, A. Rieckmann, E. C. Mormino, K. A. Johnson, R. A. Sperling, and R. L. Buckner, "Multiple brain markers are linked to age-related variation in cognition," *Cerebral cortex*, vol. 26, no. 4, pp. 1388–1400, 2016.

[3] B. of Aging Consortium, C. M. Herzog, L. J. Goeminne, J. R. Poganik, N. Barzilai, D. W. Belsky, J. Betts-LaCroix, B. H. Chen, M. Chen, A. A. Cohen *et al.*, "Challenges and recommendations for the translation of biomarkers of aging," *Nature Aging*, vol. 4, no. 10, pp. 1372–1383, 2024.

[4] A. Leuzy, N. Mattsson-Carlgren, S. Palmqvist, S. Janelidze, J. L. Dage, and O. Hansson, "Blood-based biomarkers for alzheimer's disease," *EMBO molecular medicine*, vol. 14, no. 1, p. e14408, 2022.

[5] B. Dubois, H. Hampel, H. H. Feldman, P. Scheltens, P. Aisen, S. Andrieu, H. Bakardjian, H. Benali, L. Bertram, K. Blennow *et al.*, "Preclinical alzheimer's disease: definition, natural history, and diagnostic criteria," *Alzheimer's & Dementia*, vol. 12, no. 3, pp. 292–323, 2016.

[6] B. Dubois, H. H. Feldman, C. Jacova, H. Hampel, J. L. Molinuevo, K. Blennow, S. T. DeKosky, S. Gauthier, D. Selkoe, R. Bateman *et al.*, "Advancing research diagnostic criteria for alzheimer's disease: the iwg-2 criteria," *The Lancet Neurology*, vol. 13, no. 6, pp. 614–629, 2014.

[7] E. C. Holston, "An integrative review about electrophysiological biomarkers of cognitive impairment in alzheimer's disease: a developing relationship," *Issues in Mental Health Nursing*, vol. 45, no. 7, pp. 746–757, 2024.

[8] S. A. Seyyed Mousavi, S. Javadzadeh, M. Asadi, R. B. Ivry, and T. D. Sanger, "Behaviorally coupled oscillations reveal distinct timing roles of basal ganglia and cerebellothalamic circuits in dystonia," *Dystonia*, vol. Volume 4 - 2025, 2025. [Online]. Available: https://www.frontierspartnerships.org/journals/dystonia/articles/10.3389/dyst.2025.15390

[9] J. A. Deal, J. Betz, K. Yaffe, T. Harris, E. Purchase-Helzner, S. Satterfield, S. Pratt, N. Govil, E. M. Simonsick, F. R. Lin *et al.*, "Hearing impairment and incident dementia and cognitive decline in older adults: the health abc study," *Journals of Gerontology Series A: Biomedical Sciences and Medical Sciences*, vol. 72, no. 5, pp. 703–709, 2017.

[10] G. Livingston, J. Huntley, A. Sommerlad, D. Ames, C. Ballard, S. Banerjee, C. Brayne, A. Burns, J. Cohen-Mansfield, C. Cooper *et al.*, "Dementia prevention, intervention, and care: 2020 report of the lancet commission," *The lancet*, vol. 396, no. 10248, pp. 413–446, 2020.

[11] B. S. Y. Yeo, H. J. J. M. D. Song, E. M. S. Toh, L. S. Ng, C. S. H. Ho, R. Ho, R. A. Merchant, B. K. J. Tan, and W. S. Loh, "Association of hearing aids and cochlear implants with cognitive decline and dementia: a systematic review and meta-analysis," *JAMA neurology*, vol. 80, no. 2, pp. 134–141, 2023.

[12] D. T. Gray, L. Umapathy, N. M. De La Peña, S. N. Burke, J. R. Engle, T. P. Trouard, and C. A. Barnes, "Auditory processing deficits are selectively associated with medial temporal lobe mnemonic function and white matter integrity in aging macaques," *Cerebral Cortex*, vol. 30, no. 5, pp. 2789–2803, 2020.

[13] Y. Hamza, Y. Yang, J. Vu, A. Abdelmalek, M. Malekifar, C. A. Barnes, and F.-G. Zeng, "Auditory brainstem responses as a biomarker for cognition," *Communications Biology*, vol. 7, no. 1, p. 1653, 2024.

[14] J. Majidpour, H. Hassanzadeh, E. Khezri, and H. Arabi, "Optimizing auditory brainstem response detection through nsga-ii guided feature selection," *Neural Computing and Applications*, pp. 1–19, 2025.

[15] C. Chen, L. Zhan, X. Pan, Z. Wang, X. Guo, H. Qin, F. Xiong, W. Shi, M. Shi, F. Ji *et al.*, "Automatic recognition of auditory brainstem response characteristic waveform based on bidirectional long short-term memory," *Frontiers in Medicine*, vol. 7, p. 613708, 2021.

[16] H. Wimalarathna, S. Ankmnal-Veeranna, C. Allan, S. K. Agrawal, P. Allen, J. Samarabandu, and H. M. Ladak, "Comparison of machine learning models to classify auditory brainstem responses recorded from children with auditory processing disorder," *Computer methods and programs in biomedicine*, vol. 200, p. 105942, 2021.

[17] H. Wimalarathna, S. Ankmnal-Veeranna, C. Allan, S. K. Agrawal, J. Samarabandu, H. M. Ladak, and P. Allen, "Machine learning approaches used to analyze auditory evoked responses from the human auditory brainstem: A systematic review," *Computer methods and programs in biomedicine*, vol. 226, p. 107118, 2022.

[18] S. Dass, M. S. Holi, and K. Soundararajan, "Classification of brainstem auditory evoked potentials using artificial neural network based on time and frequency domain features," *Journal of Clinical Engineering*, vol. 41, no. 2, pp. 72–82, 2016.

[19] S. N. A. Seha and D. Hatzinakos, "Human recognition using transient auditory evoked potentials: a preliminary study," *IET Biometrics*, vol. 7, no. 3, pp. 242–250, 2018.

[20] R. M. McKearney and R. C. MacKinnon, "Objective auditory brainstem response classification using machine learning," *International journal of audiology*, vol. 58, no. 4, pp. 224–230, 2019.

[21] J. Ma, J.-H. Seo, I. J. Moon, M. K. Park, J. B. Lee, H. Kim, J. H. Ahn, J. H. Jang, J. D. Lee, S. J. Choi *et al.*, "Auditory brainstem response data preprocessing method for the automatic classification of hearing loss patients," *Diagnostics*, vol. 13, no. 23, p. 3538, 2023.

[22] S. Liang, J. Xu, H. Liu, R. Liang, Z. Guo, M. Lu, S. Liu, J. Gao, Z. Ye, and H. Yi, "Automatic recognition of auditory brainstem response waveforms using a deep learning-based framework," *Otolaryngology–Head and Neck Surgery*, vol. 171, no. 4, pp. 1165–1171, 2024.

[23] M. Hashemi, "Enlarging smaller images before inputting into convolutional neural network: zero-padding vs. interpolation," *Journal of Big Data*, vol. 6, no. 1, pp. 1–13, 2019.

[24] Y. Bengio, I. Goodfellow, A. Courville *et al.*, *Deep learning*. MIT press Cambridge, MA, USA, 2017, vol. 1.

[25] S. Kiranyaz, O. Avci, O. Abdeljaber, T. Ince, M. Gabbouj, and D. J. Inman, "1d convolutional neural networks and applications: A survey," *Mechanical systems and signal processing*, vol. 151, p. 107398, 2021.

[26] B. Zhao, H. Lu, S. Chen, J. Liu, and D. Wu, "Convolutional neural networks for time series classification," *Journal of systems engineering and electronics*, vol. 28, no. 1, pp. 162–169, 2017.

[27] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[28] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921–2929.

[29] D. P. Kingma, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

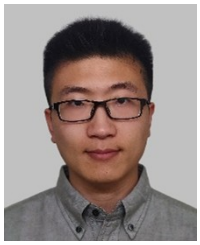
[30] M. W. Browne, "Cross-validation methods," *Journal of mathematical psychology*, vol. 44, no. 1, pp. 108–132, 2000.

[31] J. J. Eggermont, "Auditory brainstem response," *Handbook of clinical neurology*, vol. 160, pp. 451–464, 2019.

[32] J. E. Peelle and A. Wingfield, "The neural consequences of age-related hearing loss," *Trends in neurosciences*, vol. 39, no. 7, pp. 486–497, 2016.
[33] R. A. Bauder, T. M. Khoshgoftaar, and T. Hasanin, "An empirical study on class rarity in big data," in *2018 17th IEEE international conference on machine learning and applications (ICMLA)*. IEEE, 2018, pp. 785–790.
[34] S. Kerneis, E. Caillaud, and D. Bakhos, "Auditory brainstem response: Key parameters for good-quality recording," *European Annals of Otorhinolaryngology, Head and Neck Diseases*, vol. 140, no. 4, pp. 181–185, 2023.
[35] M. Asadi, S. Javadzadeh, R. Soroushmojdehi, S. A. S. Mousavi, and T. Sanger, "BACE: Behavior-adaptive connectivity estimation for interpretable graphs of neural dynamics," *bioRxiv*, 2025, also available as arXiv:2510.20831. [Online]. Available: https://www.biorxiv.org/content/early/2025/10/23/2025.10.21.683776
[36] L. Alzubaidi, J. Bai, A. Al-Sabaawi, J. Santamaría, A. S. Albahri, B. S. N. Al-Dabbagh, M. A. Fadhel, M. Manoufali, J. Zhang, A. H. Al-Timemy *et al.*, "A survey on deep learning tools dealing with data scarcity: definitions, challenges, solutions, tips, and applications," *Journal of Big Data*, vol. 10, no. 1, p. 46, 2023.
[37] R. Li, M. Hu, R. Gao, L. Wang, P. N. Suganthan, and O. Sourina, "Tformer: A time–frequency transformer with batch normalization for driver fatigue recognition," *Advanced Engineering Informatics*, vol. 62, p. 102575, 2024.

Mobina Malekifar (Graduate Student Member, IEEE) received a B.S. degree in Biomedical Engineering from the University of Isfahan, Isfahan, Iran, in 2020, and an M.S. degree in Electrical Engineering from the University of California, Irvine, in 2025. She has been involved in research on myoelectric prosthesis control, therapeutic game development, and biosignal processing. She is currently pursuing a Ph.D. degree in Electrical Engineering at the University of California, Irvine, CA, USA. Her research interests include biosignal processing and bioelectric devices.

Yasmeen Hamza received the B.S. degree in Medicine and Surgery (MBBCH) from Alexandria University, Egypt, in 2008, the Professional Postgraduate Diploma in Total Quality Management for Healthcare Reform from the American University in Cairo (AUC) in 2011, the M.Sc. degree in Audiology-Otorhinolaryngology from Alexandria University in 2013, and the Ph.D. degree in Experimental Otorhinolaryngology from KU Leuven, Belgium, in 2018. She completed postdoctoral training in Biomedical Engineering at the University of Rochester in 2019 and in Otolaryngology at the University of California, Irvine, from 2019 to 2022. In 2022, Dr. Hamza joined Florida State University as an Assistant Professor, Affiliate Faculty at the Institute for Successful Longevity, and Director of the Hearing, Speech, and Cognition Laboratory (HeSp-Cog). In 2024, she was appointed Lecturer in Audiology at the University of Southampton, United Kingdom, and became a Fellow of the Higher Education Academy (FHEA) in 2025. Her research interests focus on the integration of clinical expertise with interdisciplinary approaches to investigate the intersection of hearing disorders and cognitive function.



Ye Yang received the B.S. degree in Biology from the University of Chinese Academy of Sciences, in 2018, and currently a Ph.D. candidate in University of California, Irvine. He has been with the Hearing and Speech lab, UCI, since 2019. He is currently working on advancing audio processing techniques with artificial intelligence to help listeners with hearing loss and developing new non-invasive speech objective intelligibility metrics.

Hung Cao (Senior Member, IEEE) received his B.Sc. from Hanoi University of Science and Technology, Vietnam. He got his M.Sc. and Ph.D. in Electrical Engineering from UT Arlington in 2007 and 2012, respectively. Dr. Cao received additional training in bioengineering and medicine at University of Southern California (2012-2013) and University of California, Los Angeles (2013-2014). In 2014-2015, he worked for ETS, Montreal, QC, Canada as a research faculty. From 2015-2018, Dr. Cao worked as an Assistant Professor of Electrical and Biomedical Engineering at University of Washington, Bothell campus. Dr. Cao joined UC Irvine in 2018 and he is currently an Associate Professor of Electrical Engineering, Biomedical Engineering and Computer Science. Dr. Cao is a recipient of the National Science Foundation (NSF) CAREER Award in 2017.

Fan-Gang Zeng (S'88-M'91-SM'07-F'11) received B.S. from University of Science and Technology of China in 1982, M.S. from Academia Sinica in 1985, and Ph.D. from Syracuse University in 1990. He worked at House Ear Institute (1990-1998), University of Southern California (1996-1998), and University of Maryland, College Park (1998-2000). He is Director of Center for Hearing Research, Research Director of Otolaryngology - Head and Neck Surgery, and Chancellor's Professor of Anatomy and Neurobiology, Biomedical Engineering, and Cognitive Sciences at the University of California Irvine. Dr. Zeng has published widely in areas from engineering to medicine and public policy, including a volume on cochlear implants in Springer Handbook of Auditory Research (Springer-Verlag, New York). He has been on the editorial board for IEEE Transactions on Biomedical Engineering, Journal of Association for Research in Otolaryngology, and Hearing Research. Dr. Zeng was elected to the USA National Academy of Engineering in 2023. He has consulted for NIH, NSF, DOD, National Natural Science Foundation of China, Natural Sciences and Engineering Research Council of Canada, and numerous other public and private agencies. He is Fellow of IEEE, Acoustical Society of America, and The American Institute for Medical and Biological Engineering, and served on the Advisory Board for Apherma Corporation, Sunnyvale, CA, ImThera Medical, San Diego, CA, Nurotron Biotech Inc., Hangzhou, China, SoundCure, Boston, MA, and Syntiant, Irvine, CA.