electronics

MDPI

Article

A Deep Learning Model for Wave V Peak Detection in Auditory Brainstem Response Data

Jun Ma [1], Nak-Jun Sung [2], Sungjun Choi [3], Min Hong [4] and Sungyeup Kim [5,*]

[1] Department of Software Science and Engineering, Kunsan National University, Kunsan 54150, Republic of Korea; junma@kunsan.ac.kr
[2] National Cancer Center, Research Institute, Goyang 10408, Republic of Korea; njsung@ncc.re.kr
[3] Department of Otorhinolaryngology-Head and Neck Surgery, College of Medicine, Soonchunhyang University, Cheonan Hospital, Cheonan 31151, Republic of Korea; 64572@schmc.ac.kr
[4] Department of Computer Software Engineering, Soonchunhyang University, Asan 31538, Republic of Korea; mhong@sch.ac.kr
[5] Institute for Artificial Intelligence and Software, Soonchunhyang University, Asan 31538, Republic of Korea
* Correspondence: sungyeupkim@sch.ac.kr

Abstract

In this study, we propose a YOLO-based object detection algorithm for the automated and accurate identification of the fifth wave (Wave V) in auditory brainstem response (ABR) graphs. The ABR test plays a critical role in the diagnosis of hearing disorders, with the fifth wave serving as a key marker for clinical assessment. However, conventional manual detection is time-consuming and subject to variability depending on the examiner's expertise. To address these limitations, we developed a real-time detection method that utilizes a YOLO object detection model applied to ABR graph images. Prior to YOLO training, we employed a U-Net-based preprocessing algorithm to automatically remove existing annotated peaks from the ABR images, thereby generating training data suitable for peak detection. The proposed model was evaluated in terms of precision, recall, and mean average precision (mAP). The experimental results demonstrate that the YOLO-based approach achieves high detection performance across these metrics, indicating its potential as an effective tool for reliable Wave V peak localization in audiological applications.

Keywords: auditory brainstem response; ABR; deep learning; pre-processing; U-Net; YOLOv5; YOLOv10

1. Introduction

Auditory brainstem response (ABR) refers to the electrophysiological response generated during auditory signal transmission from the cochlear nerve to the brainstem and higher auditory pathways. The clinical procedure that measures these responses and visualizes them as waveform graphs for the assessment of auditory function is commonly known as an ABR test. Typically, ABR data consist of characteristic waveforms, identified as Wave I through Wave V, which serve as important indicators reflecting the physiological integrity of the auditory pathways [1–3].

Wave I originates from the distal portion of the auditory (cochlear) nerve and reflects the initial neural firing evoked by acoustic stimulation. Wave II is generated primarily by the proximal auditory nerve and the cochlear nucleus. Wave III is associated with neural activity in the superior olivary complex within the lower brainstem, while Wave IV is thought to arise from multiple brainstem nuclei, including the lateral lemniscus.

Check for updates

Academic Editor: George A. Tsihrintzis

Received: 2 December 2025
Revised: 12 January 2026
Accepted: 20 January 2026
Published: 25 January 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

Together, Waves I–IV provide critical information about neural conduction along the early auditory pathway and are commonly used to assess retrocochlear integrity and neural transmission abnormalities.

Among these components, Wave V, which is primarily generated in the lateral lemniscus and inferior colliculus, is the most robust and clinically reliable peak. It is widely used for estimating hearing thresholds and neural conduction latency, making it the primary target for automated detection in this study.

Among these waveforms, the Wave V peak is regarded as the most clinically significant indicator, playing a critical role in evaluating hearing thresholds and diagnosing auditory nerve disorders [4–6]. However, in practical clinical environments, ABR signals are often contaminated by noise originating from patient movement, physiological artifacts, and environmental interference, leading to variability in both the morphology and timing of the peaks [7]. Consequently, the conventional approach of manually identifying the Wave V peak by trained professionals is time-consuming and may yield inconsistent results depending on the examiner's level of expertise.

To overcome these challenges, recent studies have increasingly explored automated analysis methods incorporating artificial intelligence (AI) technologies. In particular, deep neural network-based approaches have shown remarkable performance in medical image analysis and signal processing, and have been progressively adopted for ABR signal interpretation [8–12]. Nevertheless, most existing research has focused on relatively coarse tasks, such as noise reduction and basic peak classification, and remains limited in handling more fine-grained objectives, including the precise localization of waveform peaks. Therefore, this study proposes an advanced image-based deep learning framework designed to achieve more accurate and efficient detection of the Wave V peak.

In this study, we developed a novel two-stage deep learning framework for robust Wave V peak detection. In the first stage, a U-Net model is employed to perform pixel-level preprocessing, selectively removing the existing Wave V peaks from the ABR waveform images. This procedure produces clean, peak-suppressed background images, thereby enhancing the reliability of subsequent detection. In the second stage, a You Only Look Once (YOLO)-based object detection model is trained on the preprocessed data to accurately localize the Wave V peaks within ABR graph images. By integrating the fine-grained feature removal capability of U-Net with the fast and precise localization strength of YOLO, the proposed framework significantly improves both the accuracy and efficiency of ABR signal analysis.

The primary contributions of this study are summarized as follows:

- We introduce a U-Net-based pixel-level preprocessing technique for ABR data, which enhances the accuracy of Wave V peak detection by selectively removing the target peak region.
- We propose an efficient real-time detection approach for precise Wave V peak localization through a YOLO-based object detection model.

The remainder of this paper is structured as follows: In Section 2, we review relevant studies and deep learning-based approaches for ABR analysis, followed by a detailed explanation of the proposed U-Net-based preprocessing and YOLO-based detection models. Subsequently, Section 3 presents experimental outcomes and comparative analyses with existing approaches. Finally, Sections 4 and 5 outline the contributions of this work and discuss potential directions for future research.

2. Materials and Methods

2.1. Prior Research on ABR Peak Analysis and Deep Learning Techniques

Traditional auditory brainstem response (ABR) analysis has largely depended on the visual examination performed by experienced clinicians. Although widely practiced, this approach requires substantial expertise and suffers from limited reliability and reproducibility due to its inherent subjectivity. To mitigate these limitations, early research predominantly focused on applying digital signal processing methods to reduce noise in ABR signals and to achieve automated peak detection [13–15].

2.1.1. Expert-Led Visual Waveform Algorithms

The most conventional and widely practiced method for ABR interpretation relies on experienced clinicians visually examining the recorded waveforms to derive clinical judgments.

(1) Key Analytical Components: The primary objective of this approach is to identify clinically meaningful peaks, most notably Wave I, Wave III, and Wave V.
(2) Interpretation Criteria: Clinicians typically assess the latency and amplitude of each peak, along with interpeak latencies, to infer neural conduction characteristics. Hearing thresholds, for example, are estimated by progressively decreasing the stimulus intensity and determining the lowest level at which Wave V remains observable.
(3) Inherent Limitations: Despite its widespread use, this expert-driven visual analysis is highly dependent on examiner expertise and subjective judgment. These factors, coupled with the instability of low–signal-to-noise ratio (SNR) waveforms, contribute to reduced objectivity and limited reproducibility across examiners.

2.1.2. Quantitative Signal Processing and Automated Algorithms

To overcome the inherent subjectivity of visual interpretation and to improve the objectivity of ABR analysis, a number of computational signal-processing algorithms have been proposed.

(1) Template Matching: This method detects characteristic peaks by comparing the input waveform with a predefined standard ABR template and evaluating their similarity. However, its robustness is limited due to substantial morphological variability across individuals, which often results in decreased generalization performance [16].
(2) Cross-Correlation Analysis: In this approach, the waveform obtained from a high-intensity stimulus—where the neural response is typically clearer—is used as a reference signal. The cross-correlation between this reference and the ambiguous waveforms recorded at lower stimulus intensities is computed, and a high correlation coefficient is interpreted as evidence supporting the presence of a meaningful auditory response [17].
(3) F-statistic at a Single Point (FSP): FSP is a statistical hypothesis-testing method that compares the variance of the signal and noise at a specific time point. By evaluating the presence or absence of a response based on statistical significance, FSP improves the objectivity and reliability of ABR interpretation, reducing dependence on subjective visual judgment [18].

2.1.3. Application of Classical Machine Learning Approaches

Prior to the emergence of the deep learning paradigm, several studies attempted to apply classical machine learning techniques to ABR waveform analysis.

(1) Feature Engineering–based Classification: Hand-crafted features—such as latency, amplitude, and spectral energy—were extracted from the ABR waveform. These

predefined feature vectors were then used as inputs to traditional classifiers, including support vector machines (SVMs) and artificial neural networks (ANNs), to distinguish normal from abnormal responses or to determine the presence or absence of an auditory response [19–21].

(2) Technical Limitations: In these approaches, the feature extraction step played a decisive role in overall model performance, resulting in a strong dependence on domain knowledge. This stands in stark contrast to deep learning models, which inherently learn discriminative representations directly from data without requiring manually engineered features.

In summary, pre-deep learning ABR analysis methodologies evolved from expert-driven qualitative assessment toward more objective, statistically and algorithmically grounded techniques. However, the absence of automated feature learning and the limited capacity to model complex nonlinear patterns highlighted fundamental constraints. Deep learning-based methods have since emerged as a transformative paradigm capable of addressing these long-standing challenges.

2.2. Deep Learning-Based ABR Signal Processing

Recent advances in medical signal processing have been driven by the rapid development of deep learning techniques, particularly convolutional neural networks (CNNs) and recurrent neural networks (RNNs). In ABR analysis, CNN-based approaches have been widely explored, with many studies converting ABR signals into image or spectrogram representations to detect peaks or perform classification tasks using CNN models [22–25]. Research in deep learning-based ABR processing generally falls into three major categories, as described below.

2.2.1. Automatic Detection of Major Waveform Peaks

One of the most critical steps in ABR interpretation is the accurate identification of Wave V, which holds the highest diagnostic value. Deep learning models have been effectively employed to automatically detect the position of Wave V by learning discriminative morphological characteristics of the ABR waveform.

(1) CNN-based Approaches: CNNs, known for their powerful feature extraction capabilities in image processing, have been adapted to treat ABR waveforms as one-dimensional images. Numerous studies have trained CNN models to learn the characteristic patterns of Wave V by feeding preprocessed ABR waveforms into the network. The model outputs either the probability that Wave V appears at a specific latency point or a direct prediction of its precise temporal position. These approaches have demonstrated expert-level accuracy in Wave V detection.

(2) RNN-based Approaches: Since ABR waveforms represent sequential data with inherent temporal dependencies, recurrent neural network architectures such as LSTM and GRU have also been applied. By modeling temporal context and long-range dependencies within the waveform, RNN-based methods exhibit strong robustness, particularly in noisy recording conditions, enabling reliable peak detection even when the signal quality is degraded.

2.2.2. Automatic Estimation of Hearing Threshold

One of the primary objectives of ABR testing is to determine the hearing threshold—defined as the minimum intensity level at which an auditory response is detectable. Deep learning models have been applied to learn the intensity-dependent variations in ABR waveforms and to automatically estimate this threshold.

(1) Classification-based Approaches: Several studies have developed deep learning models that classify an ABR waveform recorded at a given intensity as either "response present" or "no response." By repeating this classification across multiple intensity levels, the model identifies the lowest intensity at which a response is detected, thereby determining the hearing threshold automatically.
(2) Regression-Based Approaches: Other studies have employed regression models that take a series of ABR waveforms across intensities as input and directly predict the numeric value of the hearing threshold. This approach considers the overall pattern of waveform changes, enabling a more integrated and potentially more efficient estimation process.

2.2.3. Classification of Hearing Loss Types and Diagnostic Support

The morphology of ABR waveforms can vary depending on the type of hearing loss (e.g., conductive or sensorineural) or the presence of specific neurological disorders. Deep learning models have been explored as diagnostic decision-support tools capable of learning such subtle waveform differences.

Waveform-based Classification: In these studies, ABR waveforms from normal-hearing individuals and from patients with specific types of hearing loss are used to train deep learning classifiers. When presented with a new waveform, the model predicts the most likely diagnostic category. These approaches demonstrate the potential of deep learning to provide objective indicators derived from complex waveform patterns, thereby supporting clinical decision-making.

2.3. U-Net–Based Pixel-Level Preprocessing Techniques

With recent advances in deep learning, there has been growing interest in image processing methods that enable highly precise pixel-level manipulation. Within this context, the U-Net architecture has emerged as a key model for preprocessing tasks that require fine-grained pixel-wise transformation or restoration. Originally proposed by Ronneberger et al. [26], U-Net is a deep convolutional neural network designed for medical image segmentation. Its architectural characteristics make it particularly well-suited for tasks demanding pixel-level accuracy. Two structural components are central to its effectiveness:

(1) Symmetric Encoder–Decoder Architecture: The encoder extracts hierarchical features from the input image, compressing global contextual information, whereas the decoder progressively upsamples the compressed feature maps to reconstruct an output at the original resolution. This structure facilitates the integration of global semantic information with spatial localization cues, enabling precise reconstruction.
(2) Skip Connections: A defining characteristic of U-Net is the use of skip connections that directly transfer feature maps from each encoder layer to its corresponding decoder layer. These connections preserve low-level spatial information that might otherwise be lost during upsampling, allowing the decoder to recover fine structural details. By combining high-level semantic features with low-level positional information, U-Net achieves highly accurate pixel-level predictions.

These architectural strengths have enabled U-Net to be successfully applied to a wide range of medical imaging tasks, including the precise delineation of pathological regions and the effective removal of noise or other irrelevant structures [27]. More recently, the utility of U-Net has expanded beyond medical imaging to include various one-dimensional signal processing applications. In this study, we leverage the strengths of U-Net to perform, for the first time, a pixel-level preprocessing task that removes the Wave V peak from ABR data with high precision.

2.4. YOLO-Based Object Detection Research

Advances in deep learning have driven major breakthroughs in medical imaging, enabling earlier disease detection, quantitative lesion analysis, and improved treatment planning [28]. In particular, object detection algorithms based on Convolutional Neural Networks (CNNs) have played a central role in automatically identifying and localizing anatomical structures or abnormal lesions across a wide range of medical imaging modalities, including computed tomography (CT), magnetic resonance imaging (MRI), ultrasound, and digital pathology slides [29–31].

Early deep learning-based object detection models were dominated by the two-stage R-CNN family. Models such as R-CNN, Fast R-CNN, and Faster R-CNN achieved high detection accuracy by first proposing candidate regions and subsequently classifying them. However, their multi-stage pipelines resulted in relatively slow inference speeds, limiting their applicability in clinical environments where real-time processing is essential.

To address these limitations, the YOLO (You Only Look Once) framework was introduced as a one-stage detector that reformulates object detection as a single regression problem. YOLO partitions the input image into a grid, and each grid cell directly predicts bounding boxes and class probabilities. This novel approach dramatically improves processing speed compared with earlier models, making real-time object detection feasible.

The YOLO family has continued to evolve, with successive versions achieving significant improvements in both speed and accuracy. YOLOv2 (YOLO9000) addressed key limitations of YOLOv1, such as poor small-object detection, while YOLOv3 introduced multi-scale prediction and a deeper backbone network (Darknet-53). More recent versions—including YOLOv4, YOLOv5, and beyond—have further advanced performance through optimized anchor-box designs, improved loss functions, and refined backbone architectures. These developments have contributed to strong performance in medical imaging tasks that require high precision and robustness [32–34].

The speed and accuracy of YOLO models have enabled successful applications across numerous medical domains:

(1) Tumor and Lesion Detection: YOLO has been applied to detect pulmonary nodules, brain tumors, liver cancers, and other malignant lesions in CT and MRI scans, assisting radiologists by rapidly identifying lesion locations and reducing interpretation time [35,36].
(2) Cellular and Tissue Analysis: In digital pathology, YOLO has been used to count cancer cells or identify specific cell types on whole-slide images, contributing to increased automation and objectivity in pathological diagnosis [37,38].
(3) Medical Instrument and Anatomical Structure Identification: YOLO-based methods have been used in surgical videos or X-ray images to track surgical tools or localize bones and organs, supporting intraoperative guidance and anatomical analysis [39,40].

Given these strengths, YOLO-based object detection has become a key technology for enhancing both efficiency and accuracy in medical image analysis.

In this study, we apply the efficient object detection capabilities of YOLO to ABR signals for the first time, enabling accurate localization of Wave V peaks. The novelty of our approach lies in the integrated use of U-Net and YOLO, combining pixel-level precision in preprocessing with robust object detection for temporal localization. This unified framework addresses limitations of previous methods and establishes a new direction for automated ABR analysis.

2.5. Overall Framework Overview

The proposed method consists of a two-stage deep learning framework. The first stage performs U-Net–based preprocessing, in which the Wave V peak is removed from the ABR graph, leaving only the background signal. The second stage applies the YOLO model to the preprocessed images to detect the precise location of the peak using object detection techniques.

To ensure the integrity of this two-stage training process and prevent data leakage, we implemented a strict subject-level data partitioning strategy. The total dataset, comprising ABR recordings from 3114 subjects, was divided into two disjoint subsets. The first subset, consisting of 114 subjects, was dedicated to training and validating the U-Net model (91 training, 23 validation). The second subset, comprising the remaining 3000 subjects, was used exclusively for the YOLO model (2400 training, 600 validation). This rigorous separation guarantees that the models were trained and evaluated on completely independent patient groups.

2.5.1. Data Preprocessing (U-Net Application)

Most existing ABR datasets used for training typically consist of waveforms where the Wave V peaks have already been annotated by audiologists or clinicians. However, utilizing these annotated images directly for training poses a critical issue: the object detection model tends to learn the visual features of the 'marker' itself rather than the morphological characteristics of the Wave V peak. In a real-world clinical setting, the primary goal of this AI framework is to assist clinicians by detecting peaks on raw, unannotated ABR data before manual review. Therefore, to simulate this pre-diagnosis environment and prevent the model from relying on pre-existing markers, the U-Net preprocessing step is essential. It artificially removes the annotated peaks from the training data, restoring the waveform to its 'raw' state, which forces the YOLO model to learn the intrinsic features of the Wave V peak.

To selectively remove the Wave V peak from ABR signal images, this study adopts a simplified 2D U-Net architecture as the segmentation network. The 2D U-Net, structured in an encoder–decoder configuration, extracts key patterns from the input image through the encoding path and removes peak-related pixels during the decoding process without restoring them. Skip connections are incorporated between corresponding encoder and decoder layers to minimize the loss of fine-grained details by transferring adjacent feature maps.

Figure 1 illustrates the simple 2D U-Net architecture employed in this study.

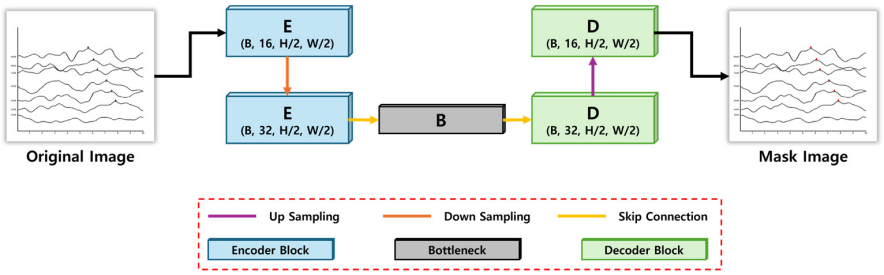


Figure 1. A Simple 2D U-Net architecture proposed in this study.

The model was trained using the proposed architecture, and the major hyperparameters used for training were as follows:

1. Optimizer: Adam;
2. Learning rate: 1×10^{-4};
3. Batch size: 8;

4. Loss function: Binary Cross-Entropy Loss (BCE Loss);
5. Epochs: 100.

To improve generalization performance and enhance robustness, various image augmentation techniques were applied, including random horizontal/vertical flips, rotations, and brightness adjustment. The dataset was split into training and validation sets at an 8:2 ratio. The number of samples used for training and validation was 91 and 23, respectively.

The input to the model consisted of ABR images containing the Wave V peak, accompanied by corresponding masks that indicated the peak region and served as supervisory signals. The model output was a peak-removed image, and training was performed by minimizing the pixel-wise differences between the predicted images and the ground-truth peak-removed images using the BCE loss function.

To identify the best-performing epoch during training, the following evaluation metrics were used:

1. Mean Squared Error (MSE);
2. Peak Signal-to-Noise Ratio (PSNR);
3. Structural Similarity Index Measure (SSIM).

The model demonstrated rapid convergence within the first 10 epochs. As shown in Figure 2a, the MSE stabilized at a negligible value (0.007), indicating precise pixel-level reconstruction. Simultaneously, the PSNR (Figure 2b) reached a high fidelity of approximately 21.5 dB. Most importantly, the SSIM (Figure 2c), which evaluates the structural integrity of the waveform, improved dramatically from the baseline values to a value exceeding 0.90. The inset plots highlight the training behavior during the final stabilization phase (Epochs 80–100). The minimal divergence between the training and validation curves across all three metrics confirms that the model has learned robust features of the ABR background signals without overfitting to the training data.

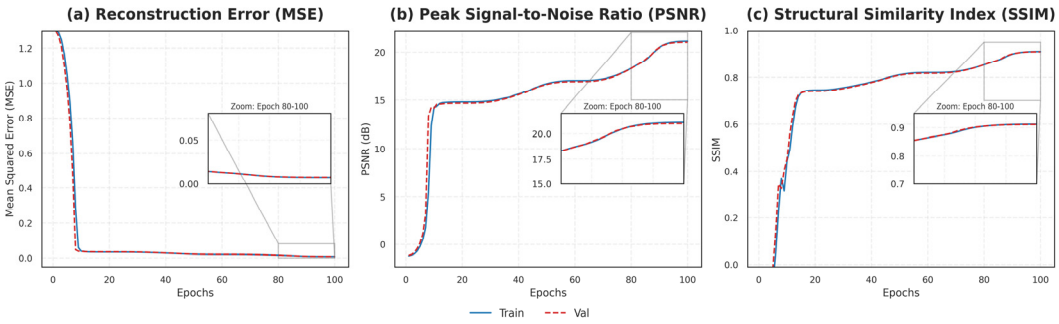


Figure 2. Comprehensive evaluation of U-Net training stability and reconstruction quality: (**a**) Mean Squared Error (MSE), (**b**) Peak Signal-to-Noise Ratio (PSNR), and (**c**) Structural Similarity Index Measure (SSIM) over 100 epochs.

Visual Validation of Peak Removal: To complement the quantitative metrics, we performed a qualitative visual inspection on the validation dataset to assess the practical utility of the preprocessing model. Figure 3 presents representative examples of ABR waveforms before and after U-Net processing.

As observed in the comparison, the model successfully identified and suppressed the target Wave V peaks (indicated by vertical markers in the original images) while preserving the complex morphological characteristics of the background signal. The processed waveforms exhibit smooth continuity without introducing artifacts or significant distortion to the surrounding peaks (e.g., Wave I, III). This visual confirmation corroborates the high SSIM scores obtained during training, demonstrating that the U-Net acts as an effective, content-aware filter that isolates the background signal for the subsequent detection stage.

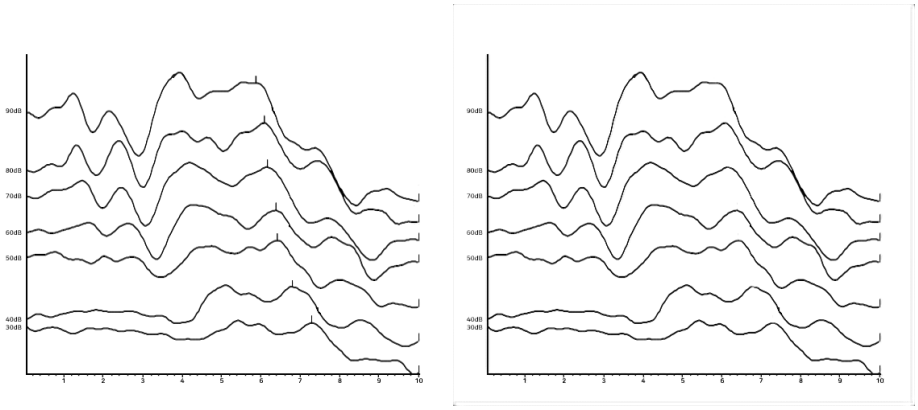


Figure 3. Peak removal results using the proposed 2D U-Net.

2.5.2. Wave V Peak Detection (YOLO Application)

Following the U-Net preprocessing, which successfully removed the visual artifacts of prior clinical annotations, we proceeded to prepare the training dataset for the object detection model. It is crucial to note that while the visual markers were removed, the coordinate information of the true Wave V peak was preserved as ground truth. We performed bounding box labeling on these 'cleaned' background images corresponding to the true peak locations. This step ensures that the YOLO model learns to identify the peak based solely on the morphological features of the waveform, rather than relying on visual cues from existing clinical markers. Subsequently, the testing phase was conducted on unlabeled ABR images (processed by U-Net) to evaluate the model's capability to autonomously detect Wave V in a raw data environment, simulating a real-world pre-diagnostic scenario.

The YOLO model provides fast and accurate object localization by predicting bounding boxes directly from images. In this study, the preprocessed images were used as inputs to YOLO to train the model to detect the region where the Wave V peak was originally located. Owing to YOLO's high-speed object detection capability, real-time analysis is feasible, making the approach highly suitable for clinical environments that require rapid evaluation. Among the various versions of YOLO, YOLOv5 and YOLOv10 were adopted in this study to train and compare performance on Wave V peak detection.

(1) YOLOv5 Framework: YOLOv5, implemented in the PyTorch 2.0.1 framework, offers high flexibility and productivity while achieving an efficient balance between speed and accuracy. Its architecture consists of three primary components: the backbone, neck, and head.

   A. Backbone: The backbone extracts hierarchical feature maps from the input image. YOLOv5 utilizes CSPDarknet53 (Cross Stage Partial Darknet53) as its backbone, which enhances computational efficiency and ensures stable gradient flow. This design improves both learning effectiveness and inference speed.

   B. Neck: The neck fuses multi-scale feature maps extracted by the backbone to improve multi-scale object detection performance. YOLOv5 combines the SPPF (Spatial Pyramid Pooling–Fast) module and PANet (Path Aggregation Network) for this purpose. SPPF enriches contextual information through multiple receptive fields, while PANet enhances feature representation by bidirectionally integrating high-level semantic features with low-level localized features.

   C. Head: The head predicts bounding boxes, objectness scores, and class probabilities based on the aggregated feature maps. YOLOv5 employs an anchor-based detection strategy and performs predictions at three different scales, enabling robust detection of objects of various sizes.

(2) YOLOv10 Framework: YOLOv10 is a fully end-to-end object detection model designed to eliminate inefficiencies associated with traditional post-processing stages. Its primary innovation is the NMS-free design, which removes the Non-Maximum Suppression (NMS) step to minimize inference latency.

A. Consistent Dual Assignments:

To remove the need for NMS, YOLOv10 introduces a novel dual-assignment strategy during training. It simultaneously applies a one-to-many assignment, where multiple predictions are linked to a single object, and a one-to-one assignment, where only one prediction is assigned.

The one-to-many head provides rich supervision signals during training, facilitating model optimization. The one-to-one head is trained to produce a single non-redundant prediction per object during inference, eliminating the need for NMS. At inference time, the one-to-many head is removed, enabling a truly NMS-free architecture without additional computational overhead.

B. Comprehensive Efficiency–Accuracy–Driven Design: Beyond the NMS-free structure, YOLOv10 incorporates several architectural optimizations to maximize performance:

i. Lightweight Classification Head: The classification head is simplified to reduce computational overhead by acknowledging the differing complexity requirements of classification and regression tasks, resulting in improved inference efficiency.

ii. Spatial–Channel Decoupled Downsampling: By separating spatial and channel operations during downsampling, YOLOv10 minimizes information loss while enhancing computational efficiency.

iii. Rank-Guided Block Design: Structural redundancy in each stage of the network is reduced based on a ranking mechanism, lowering parameter counts and computational load while maintaining representational capacity.

We selected YOLOv5 and YOLOv10 for specific reasons. YOLOv5 is widely recognized as a stability benchmark in industrial and medical applications, offering a proven trade-off between speed and accuracy. YOLOv10 was chosen to evaluate the latest advancements in NMS-free (Non-Maximum Suppression-free) architectures, which minimize post-processing latency—a critical factor for real-time clinical monitoring. Other models like YOLOv7 or v8 were considered but excluded to focus the comparison on a stable baseline (v5) versus a latency-optimized architecture (v10).

In this study, we constructed a Wave V peak detection model by first removing the Wave V peak using U-Net, then labeling the true peak location and region on the corresponding ABR images, and finally training the YOLO model directly with these annotated data.

For YOLO training, labeling was performed on the original Wave V peak regions that had been removed during preprocessing, as illustrated in Figure 4. These annotated images were then used to train and evaluate the YOLO model.

To enhance generalization performance and improve robustness during training, all images were resized to a uniform size and standardized to the same resolution. The dataset was divided into training and validation sets at an 8:2 ratio, consisting of the following number of samples:

1. Training: 2400;
2. Validation: 600.

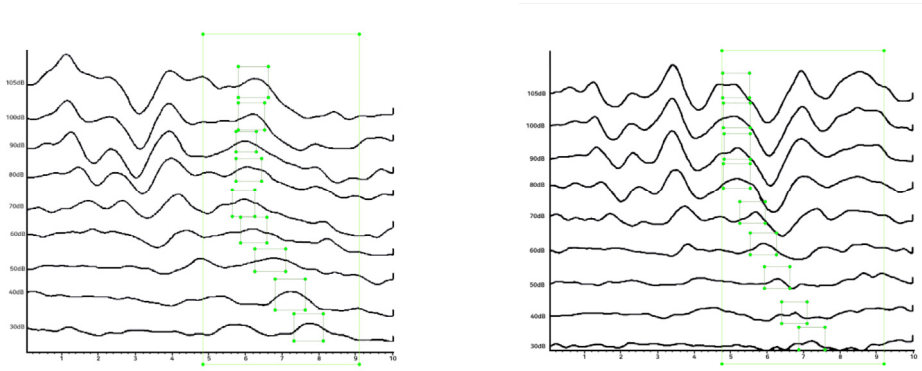


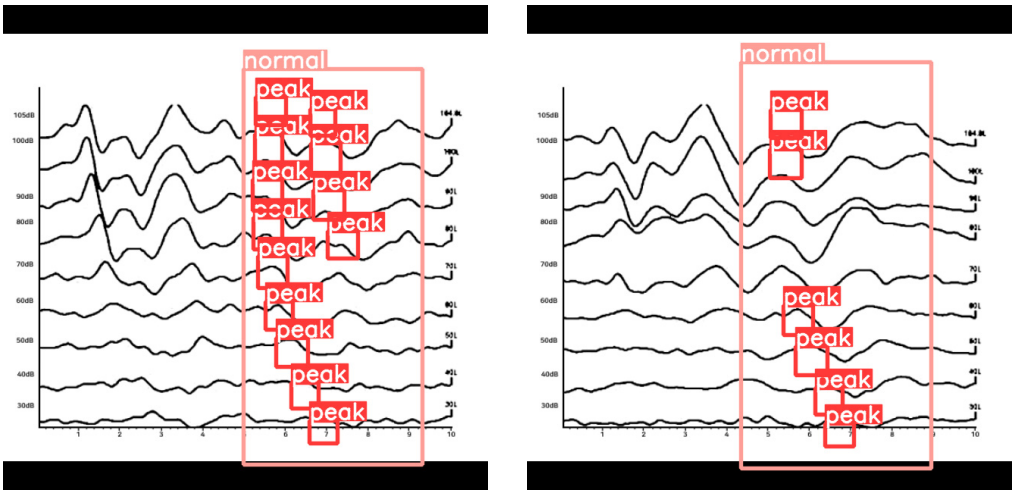
Figure 4. ABR image labeling(green box) progress screen after preprocessing.

After completing the preprocessing and labeling procedures, the peak-removed ABR images were used as inputs for training and validation, followed by evaluation on test data that were not included in the training set.

For training, we utilized the following hyperparameters for both YOLO models to ensure a fair comparison:

1. Optimizer: SGD with momentum 0.937;
2. Learning Rate: Initial lr 0.01, reducing with a cosine scheduler;
3. Batch Size: 16;
4. Epochs: 100;
5. Input Resolution: 640 × 640;
6. Box Loss Gain: 0.05, Class Loss Gain: 0.5.

As illustrated in Figure 5, the model was able to detect the Wave V peak; however, the limited size of the training dataset led to inconsistent performance, with occasional inaccuracies in distinguishing valid peak patterns. Nonetheless, preliminary experiments indicate that the detection accuracy steadily improves as the volume of training data increases, suggesting that the model possesses strong potential for robust Wave V peak localization when trained with a sufficiently large dataset. A detailed quantitative evaluation of the model's performance is presented in the following section.



(**a**) (**b**)

Figure 5. (**a**) Initial training results for YOLOv10s (train: 80, validation:20); (**b**) mid-training results for YOLOv5l (training: 800, validation: 200).

3. Results

3.1. Performance Evaluation Method

The performance of the proposed model was evaluated using several metrics, including precision, recall, and mean average precision (mAP). Table 1 presents the performance scores obtained from training the YOLOv5s, YOLOv5l, YOLOv5x, YOLOv10n, and YOLOv10x models.

Table 1. The performance scores obtained from training the YOLO models.

Model	Precision	Recall	mAP50	mAP50–95
YOLOv5s	0.885	0.758	0.865	0.785
YOLOv5l	0.891	0.887	0.895	0.811
YOLOv5x	0.911	0.941	0.944	0.891
YOLOv10n	0.891	0.801	0.884	0.799
YOLOv10x	0.888	0.911	0.911	0.798

Within the YOLOv5 family, performance generally improved as the model size increased. Notably, YOLOv5x achieved the highest scores across all metrics, with a precision of 0.911, a recall of 0.941, a mAP50 of 0.944, and a mAP50–95 of 0.891. These results suggest that the larger backbone of YOLOv5x provides richer feature representations and more stable detection capability for Wave V peak localization. For the YOLOv10 models, YOLOv10x outperformed YOLOv10n in Recall (0.911) and mAP50 (0.911), demonstrating competitive detection performance overall. However, YOLOv5x remained superior in mAP50–95, indicating stronger robustness across varying IoU thresholds. Although YOLOv10 incorporates innovative design elements—such as an NMS-free architecture—it appears that YOLOv5x achieves higher overall precision for the characteristics of the dataset used in this study. Overall, YOLOv5x delivered the best performance among all tested models, exhibiting the highest accuracy and precision for Wave V peak detection. The YOLOv10 models, while slightly less accurate in this setting, offer an efficient architecture that may provide advantages in real-time clinical applications. To verify the stability of the training process and ensure the model achieved precise localization without overfitting, we analyzed the loss curves for the YOLOv5x model. Figure 6 illustrates the progression of both classification loss and bounding box regression loss for the training and validation datasets over 100 epochs.

The top row displays the classification loss (cls_loss), while the bottom row shows the bounding box regression loss (box_loss). Both metrics demonstrate rapid convergence and stabilization, indicating that the model effectively learned to classify Wave V and precisely regress its spatial coordinates. As observed in the figure, the classification loss (top row) decreased sharply during the initial training phase and stabilized around epoch 20, confirming the model's ability to distinguish Wave V from the background. Crucially, the bounding box loss (bottom row) also showed a steady decline and convergence, which indicates that the model progressively refined the accuracy of the predicted peak coordinates. The minimal divergence between training and validation curves across both metrics validates the robustness of the trained model.

Based on these results, the YOLOv5x model—identified as the most optimal configuration—was used for inference to detect the Wave V peak by directly inputting arbitrary ABR graphs, and the corresponding outcomes are presented in Figure 7.

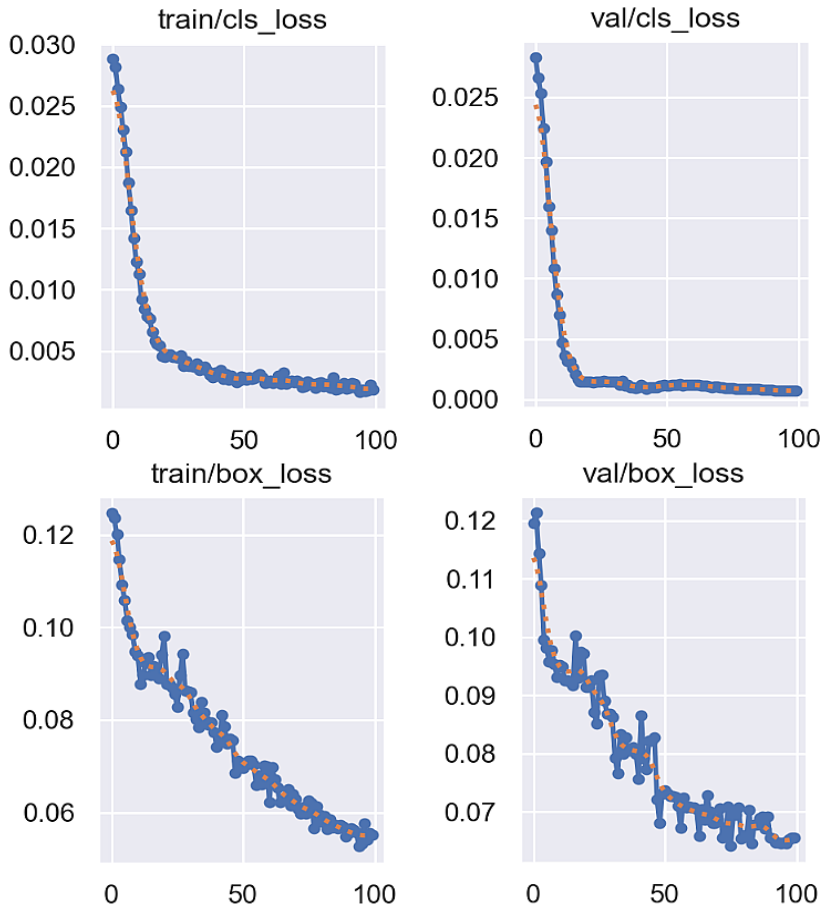


Figure 6. Training and validation loss curves for the YOLOv5x model.

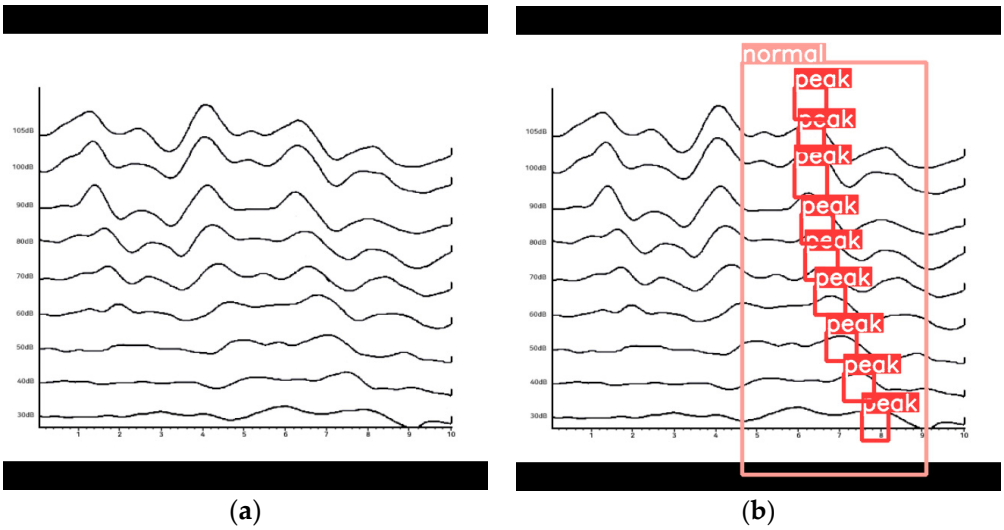


Figure 7. Following the final inference, the test results show (**a**) the ABR images used in the test and (**b**) the input to the fully trained YOLOv5x.

3.2. Computational Efficiency and Real-Time Feasibility

In addition to reconstruction performance, the computational efficiency of the preprocessing model is a critical factor for clinical deployment. We evaluated the model complexity and inference speed using a standard computing environment. As summarized in Table 2, the proposed lightweight 2D U-Net architecture contains only 116.75 K parameters, requiring significantly fewer computational resources than conventional segmentation

networks. Its computational cost is 7.11 GMACs (14.22 GFLOPs), and the average inference time for a single ABR image was 2.29 ± 0.02 ms (mean ± std, calculated over five runs), confirming that peak-removal preprocessing can be performed with negligible latency.

Table 2. Model complexity and inference performance of the proposed U-Net.

Stage	Metric	Value	Unit
U-Net (Preprocessing)	Parameters	116.75	K (Thousands)
	Computational Cost	7.11	GMACs
	Inference Time	2.29 ± 0.02	ms (Milliseconds)
YOLO (Detection)	Inference Time	5.25 ± 0.06	ms (Milliseconds)
End-To-End (U-Net + YOLO)	Total Inference Time	7.54 ± 0.06	ms (Milliseconds)

The YOLO-based Wave V detection stage was evaluated under the same hardware conditions. The average inference time required for detecting the Wave V peak in a preprocessed ABR image was 5.25 ± 0.06 ms, which demonstrates that the detection module is also computationally efficient. When both stages are combined, the complete end-to-end pipeline (U-Net preprocessing followed by YOLO detection) achieves a total inference time of approximately 7.54 ms per ABR image.

This performance comfortably satisfies real-time processing requirements for clinical ABR systems, where waveform acquisition and display typically operate at much lower temporal resolutions. Consequently, the proposed two-stage framework enables on-the-fly peak removal and Wave V localization during ABR data acquisition, supporting immediate clinical interpretation without perceptible delay.

3.3. Comparison with Another Methods

Since this study represents the first application of the YOLO framework to ABR Wave V peak detection, a direct performance comparison with identical object detection baselines is not feasible. Therefore, to validate the effectiveness of our proposed method, we compared its performance with state-of-the-art deep learning-based ABR analysis models reported in the literature. Table 3 summarizes the comparison results.

Table 3. Performance comparison with other ABR Wave V detection methods.

Study	Input Type	Architecture	Metric (Accuracy/mAP)
Liang et al. [23]	Signal	CNN + BiLSTM + Attention	91.0% (Accuracy)
Liu et al. [24]	Image	ResPatchTST	91.9% (Accuracy)
Proposed Method	Image	U-Net + YOLOv5x	94.4% (mAP50)

Although a direct comparison is limited due to differences in datasets and evaluation metrics (Accuracy vs. mAP), our proposed YOLOv5x model demonstrates competitive performance (mAP 0.944) comparable to the high accuracy (91.9%) of the ResPatchTST classification model [24]. This indicates that our framework successfully achieves precise peak localization (object detection) while maintaining a detection capability equivalent to state-of-the-art classification approaches.

4. Discussion

This study demonstrates that separating the task into peak removal (preprocessing) and location prediction (detection) yields superior performance compared with direct detection on raw ABR images. The U-Net module functions as a content-aware filter

that suppresses background components while preserving the morphology of the Wave V region, thereby improving the robustness of the subsequent YOLO-based detection stage.

Ideally, the effectiveness of the U-Net preprocessing should be validated by comparing it with a baseline model trained directly on the original annotated images. However, in our preliminary analysis, we observed that training a detection model on images containing clinical markers leads to a critical issue known as 'shortcut learning.' The model converges by simply identifying the visual markers (annotations) rather than learning the morphological features of the Wave V peak. Consequently, such a baseline model fails completely when tested on raw, unannotated data, rendering a quantitative comparison (e.g., mAP) scientifically invalid. This inherent limitation of using retrospectively annotated clinical data underscores the necessity of our proposed U-Net-based marker removal technique. While our previous studies [41,42] focused on classifying hearing status (Normal vs. Abnormal), the present work addresses a more fundamental clinical requirement: the precise localization of the Wave V peak, which is essential for estimating ABR latency and hearing thresholds.

In addition, the proposed framework satisfies real-time processing requirements. In our experiments, the U-Net preprocessing module required approximately 2 ms per ABR image, and the subsequent YOLO-based peak detection required approximately 5 ms, resulting in a total end-to-end inference time of about 7 ms per sample on a single GPU. This latency is well below the temporal constraints of clinical ABR acquisition and therefore supports real-time or near-real-time deployment in practical diagnostic settings.

A limitation of this study is that only normal-hearing data were included. Pathological ABR waveforms often exhibit distorted morphologies, such as prolonged latencies and reduced amplitudes, which may challenge both the preprocessing and detection stages. Future work will extend the dataset to include diverse hearing-impaired populations and further evaluate the generalization ability of the proposed two-stage framework under clinically more complex conditions.

5. Conclusions

Research on ABR waveform analysis using deep learning has recently emerged as a promising alternative to traditional manual procedures, effectively addressing issues related to subjectivity and inefficiency. In particular, models based on CNNs and RNNs have demonstrated strong capability in learning the spatiotemporal characteristics of ABR signals, achieving high accuracy in core clinical tasks such as peak detection, hearing threshold estimation, and hearing loss classification. Future research directions are anticipated to include the development of lightweight models capable of robust performance with limited data, the construction of large-scale datasets encompassing diverse populations, and the adoption of explainable AI (XAI) techniques to improve interpretability and clinical trust. Collectively, such advancements are expected to accelerate the automation of ABR analysis and contribute to faster and more reliable hearing diagnostic services.

Author Contributions: Conceptualization, J.M. and S.K.; methodology, J.M.; software, J.M.; validation, J.M. and S.C.; formal analysis, J.M. and M.H.; investigation, J.M., N.-J.S. and M.H.; resources, J.M., N.-J.S. and S.C.; data curation, J.M. and S.C.; writing—original draft preparation, J.M.; writing—review and editing, S.C., M.H. and S.K.; visualization, J.M.; supervision, S.K.; project administration, M.H.; funding acquisition, M.H. and S.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the BK21 FOUR (Fostering Outstanding Universities for Research), Grant No. 5199990914048, and was supported by the Soonchunhyang University Research Fund (No. 20240949) and the Kunsan National University Research Fund.

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki and approved by the Institutional Review Board of Soonchunhyang University Cheonan Hospital (Cheonan, Korea) (IRB number: 2021-06-040, approval date: 4 June 2021).

Data Availability Statement: The data presented in this study are available in a publicly accessible repository (AI Hub, https://aihub.or.kr/). Access to the dataset is subject to registration and approval in accordance with the repository's data usage policies.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Jewett, D.L.; Romano, M.N.; Williston, J.S. Human Auditory Evoked Potentials: Possible Origins as Indicated by Surface Recordings and Intracranial Records of the Cat. *Electroencephalogr. Clin. Neurophysiol.* **1970**, *28*, 278–287.
2. Don, M.; Eggermont, J.J. Analysis of the Click-Evoked Auditory Brainstem Response in Terms of Frequency-Specific Components. *J. Acoust. Soc. Am.* **1978**, *63*, 1084–1092. [CrossRef] [PubMed]
3. Hall, J.W., III. *New Handbook of Auditory Evoked Responses*; Pearson Education: Boston, MA, USA, 2007; pp. 154–196.
4. Gorga, M.P.; Johnson, T.A.; Kaminski, J.R.; Beauchaine, K.L.; Shi, Y.; Peters, J. Development of clinical norms for the click auditory brainstem response. *Ear Hear.* **2006**, *27*, 49–67.
5. Luts, H.; Dehaene, S.; Clavier, O.; Wouters, J. Estimation of hearing thresholds in children using the auditory steady-state response and the auditory brainstem response. *J. Acoust. Soc. Am.* **2006**, *120*, 3845–3852.
6. Stapells, D.R. Threshold estimation by the auditory brainstem response: Current status and future directions. *Semin. Hear.* **2010**, *31*, 12–23.
7. Hyde, M.L. Effects of the far-field recording electrode on the auditory brainstem response. *Semin. Hear.* **1990**, *11*, 320–332.
8. McKearney, R.M.; MacKinnon, R.C. Objective auditory brainstem response classification using machine learning. *Int. J. Audiol.* **2019**, *58*, 224–230. [CrossRef] [PubMed]
9. Delgado, R.E.; Ozdamar, O. Automated auditory brainstem response interpretation. *IEEE Eng. Med. Biol. Mag.* **2002**, *13*, 227–237. [CrossRef]
10. Bradley, A.P.; Wilson, W.J. Automated analysis of the auditory brainstem response. In *Proceedings of the 2004 Intelligent Sensors, Sensor Networks and Information Processing Conference*; IEEE: New York, NY, USA, 2004; pp. 541–545.
11. Shaheen, L.A.; Buran, B.N.; Suthakar, K.; Koehler, S.D.; Chung, Y. ABRpresto: An algorithm for automatic thresholding of the Auditory Brainstem Response using resampled cross-correlation across subaverages. *Hear. Res.* **2025**, *462*, 109258. [CrossRef]
12. McKearney, R.M.; Simpson, D.M.; Bell, S.L. Automated wave labelling of the auditory brainstem response using machine learning. *Int. J. Audiol.* **2025**, *64*, 766–771. [CrossRef] [PubMed]
13. McKearney, R.M.; Bell, S.L.; Chesnaye, M.A.; Simpson, D.M. Auditory Brainstem Response Detection Using Machine Learning: A Comparison with Statistical Detection Methods. *Ear Hear.* **2022**, *43*, 949–960. [CrossRef] [PubMed]
14. Majidpour, A.; Jameel, S.K.; Majidpour, J.; Bagheri, H.; Rashid, T.A.; Nazeri, A.; Aleaba, M.M. Detection of auditory brainstem response peaks using image processing techniques in infants with normal hearing sensitivity. *Biomed. Signal Process. Control* **2023**, *86*, 105117. [CrossRef]
15. Wimalarathna, H.; Ankmnal-Veeranna, S.; Allan, C.; Agrawal, S.K.; Allen, P.; Samarabandu, J.; Ladak, H.M. Comparison of machine learning models to classify auditory brainstem responses recorded from children with auditory processing disorder. *Comput. Methods Programs Biomed.* **2021**, *200*, 105942. [CrossRef]
16. Liang, S.; Xu, J.; Liu, H.; Liang, R.; Guo, Z.; Lu, M.; Liu, S.; Gao, J.; Ye, Z.; Yi, H. Automatic Recognition of Auditory Brainstem Response Waveforms Using a Deep Learning-Based Framework. *Otolaryngol.–Head Neck Surg.* **2024**, *171*, 1165–1171. [CrossRef] [PubMed]
17. Chen, C.; Zhan, L.; Pan, X.; Wang, Z.; Guo, X.; Qin, H.; Xiong, F.; Shi, W.; Shi, M.; Ji, F.; et al. Automatic recognition of auditory brainstem response characteristic waveform based on bidirectional long short-term memory. *Front. Med.* **2021**, *7*, 613708. [CrossRef] [PubMed]
18. Wimalarathna, H.; Ankmnal-Veeranna, S.; Allan, C.; Agrawal, S.K.; Samarabandu, J.; Ladak, H.M.; Allen, P. Machine learning approaches used to analyze auditory evoked responses from the human auditory brainstem: A systematic review. *Comput. Methods Programs Biomed.* **2022**, *226*, 107118. [CrossRef]

19. Dobrowolski, A.; Suchocki, M.; Tomczykiewicz, K.; Majda-Zdancewicz, E. Classification of auditory brainstem response using wavelet decomposition and SVM network. *Biocybern. Biomed. Eng.* **2016**, *36*, 427–436. [CrossRef]
20. Al Osman, R.; Al Osman, H. On the use of machine learning for classifying auditory brainstem responses: A scoping review. *IEEE Access* **2021**, *9*, 110592–110600. [CrossRef]
21. Dass, S.; Holi, M.S.; Soundararajan, K. Classification of brainstem auditory evoked potentials using artificial neural network based on time and frequency domain features. *J. Clin. Eng.* **2016**, *41*, 72–82. [CrossRef]
22. Özyurt, F.; Majidpour, J.; Rashid, T.A.; Majidpour, A.; Koc, C. Multi-transfer learning techniques for detecting auditory brainstem response. *Appl. Acoust.* **2023**, *212*, 109604. [CrossRef]
23. Liu, Y.; Xiang, L.; Li, Q.; Li, K.; Yang, Y.; Wang, T.; Qin, Y.; Fu, X.; Zhao, Y.; Gao, C. Comparison of Deep Learning Models for Objective Auditory Brainstem Response Detection: A Multicenter Validation Study. *Trends Hear.* **2025**, *29*, 23312165251347773. [CrossRef] [PubMed]
24. Baby, D.; Van Den Broucke, A.; Verhulst, S. A convolutional neural-network model of human cochlear mechanics and filter tuning for real-time applications. *Nat. Mach. Intell.* **2021**, *3*, 134–143. [CrossRef] [PubMed]
25. Drakopoulos, F.; Baby, D.; Verhulst, S. A convolutional neural-network framework for modelling auditory sensory cells and synapses. *Commun. Biol.* **2021**, *4*, 827. [CrossRef]
26. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*; Navab, N., Hornegger, J., Wells, W., Frangi, A., Eds.; Springer: Cham, Switzerland, 2015; Volume 9351, pp. 234–241.
27. Fischer, P.; Ronneberger, O.; Brox, T. The U-Net architecture for deep learning in medical image segmentation. *Methods Cell Biol.* **2017**, *141*, 119–147.
28. Litjens, G.; Kooi, T.; Bejnordi, B.E.; Setio, A.A.A.; Ciompi, F.; Ghafoorian, M.; Van Der Laak, J.A.; Van Ginneken, B.; Sánchez, C.I. Deep learning in medical image analysis. *Med. Image Anal.* **2017**, *42*, 60–81. [CrossRef] [PubMed]
29. Ciresan, D.C.; Giusti, A.; Gambardella, L.M.; Schmidhuber, J. Deep neural networks segment neuronal membranes in electron microscopy images. *Adv. Neural Inf. Process. Syst. (NIPS)* **2012**, *25*, 2843–2851.
30. Esteva, A.; Kuprel, B.; Novoa, R.A.; Ko, J.; Swetter, S.M.; Blau, H.M.; Thrun, S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **2017**, *542*, 115–118. [CrossRef] [PubMed]
31. Gulshan, V.; Peng, L.; Coram, M.; Stumpe, M.C.; Wu, F.; Nazarreta, A.; Venugopalan, S.; Widner, K.; Madams, T.; Cuadros, J.; et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA* **2016**, *316*, 2402–2410. [CrossRef]
32. Jocher, G.; Chaurasia, B.; Stoken, T.; Borovec, J. Ultralytics YOLOv5. *GitHub Repository*. 2020. Available online: https://github.com/ultralytics/yolov5 (accessed on 1 December 2025).
33. Redmon, J.; Farhadi, A. YOLO9000: Better, Faster, Stronger. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Honolulu, HI, USA, 2017; pp. 6517–6525.
34. Kaur, A.; Singh, Y.; Neeru, N.; Kaur, L.; Singh, A. A survey on deep learning approaches to medical images and a systematic look up into real-time object detection. *Arch. Comput. Methods Eng.* **2022**, *29*, 2071–2111. [CrossRef]
35. Devi, P.J.; Meghana, N.; Somayajulu, P.S.; Sohitha, P. Enhancing Pulmonary Nodule Detection in 3D CT Scans with YOLOv7 for Improved Lung Cancer Diagnosis. In *2025 International Conference on Computing Technologies (ICOCT)*; IEEE: New York, NY, USA, 2025; pp. 1–8.
36. Maheshwari, G.; Vashisht, A.; Rodwal, A.; Kothari, A.; Gonge, S. Enhanced Brain Tumor Detection Using Convolutional Neural Networks and YOLO V4 on MRI Images. In *International Conference on Innovative Computing and Communication*; Springer Nature: Singapore, 2024; pp. 405–425.
37. Escobar, F.I.F.; Alipo-on, J.R.T.; Novia, J.L.U.; Tan, M.J.T.; Karim, H.A.; AlDahoul, N. Automated counting of white blood cells in thin blood smear images. *Comput. Electr. Eng.* **2023**, *108*, 108710. [CrossRef]
38. Aly, G.H.; Marey, M.; El-Sayed, S.A.; Tolba, M.F. YOLO based breast masses detection and classification in full-field digital mammograms. *Comput. Methods Programs Biomed.* **2021**, *200*, 105823. [CrossRef] [PubMed]
39. Wang, Y.; Guo, Y.; He, G.; Li, M. Detection Technology of Assisted Identification Laparoscopic Surgical Instruments Based on YOLO. In *2024 3rd International Conference on Computing, Communication, Perception and Quantum Technology (CCPQT)*; IEEE: New York, NY, USA, 2024; pp. 338–343.
40. Putra, R.H.; Astuti, E.R.; Putri, D.K.; Widiasri, M.; Laksanti, P.A.M.; Majidah, H.; Yoda, N. Automated permanent tooth detection and numbering on panoramic radiograph using a deep learning approach. *Oral Surg. Oral Med. Oral Pathol. Oral Radiol.* **2024**, *137*, 537–544. [CrossRef] [PubMed]

41. Ma, J.; Seo, J.H.; Moon, I.J.; Park, M.K.; Lee, J.B.; Kim, H.; Ahn, J.H.; Jang, J.H.; Lee, J.D.; Choi, S.J.; et al. Auditory Brainstem Response Data Preprocessing Method for the Automatic Classification of Hearing Loss Patients. *Diagnostics* **2023**, *13*, 3538. [CrossRef] [PubMed]
42. Ma, J.; Choi, S.J.; Kim, S.; Hong, M. Performance Comparison of Convolutional Neural Network-Based Hearing Loss Classification Model Using Auditory Brainstem Response Data. *Diagnostics* **2024**, *14*, 1232. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.